Convex Relaxations of ReLU Neural Networks Approximate Global Optima in Polynomial Time

Sungyoon Kim [1] **Mert Pilanci** [1]

Abstract

In this paper, we study the optimality gap between two-layer ReLU networks regularized with weight decay and their convex relaxations. We show that when the training data is random, the relative optimality gap between the original problem and its relaxation can be bounded by a factor of $O(\sqrt{\log n})$, where n is the number of training samples. A simple application leads to a tractable polynomial-time algorithm that is guaranteed to solve the original non-convex problem up to a logarithmic factor. Moreover, under mild assumptions, we show that local gradient methods converge to a point with low training loss with high probability. Our result is an exponential improvement compared to existing results and sheds new light on understanding why local gradient methods work well.

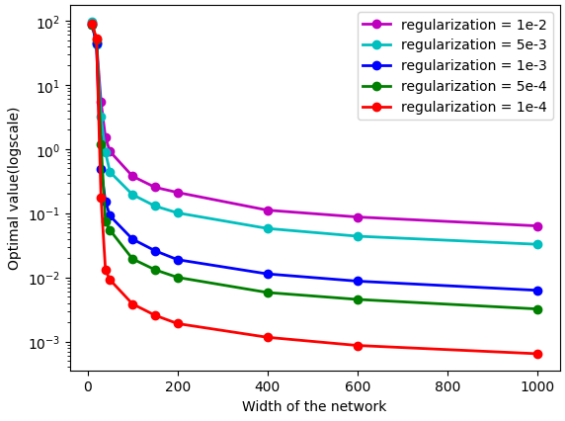

Figure 1. Convex relaxations of different widths. Here, we show the optimal value of the relaxed problem for different numbers of subsampled hyperplane arrangement patterns and regularization. Note that $n = 300, d = 10$, hence there are $\approx 30^{10}$ variables in the convex reformulation. However, using ≈ 30 neurons in the relaxed problem optimizes the objective well.

1. Introduction

After the tremendous success of deep learning (LeCun et al., 2015), data-driven approaches have become a prominent trend in various areas of computer science. Perhaps a surprising fact is that despite the highly non-convex landscape of deep learning models (Li et al., 2018), local gradient methods such as stochastic gradient descent (SGD) (Bottou, 2010) or ADAM (Kingma & Ba, 2014) find nearly-global minimizers of the network extremely well. This mystery has gained wide attention in the learning theory community, and many have worked on the problem of proving convergence results of local methods under certain assumptions such as the infinite-width limit (Jacot et al., 2018), heavy overparametrization (Du et al., 2019), (Arora et al., 2019),(Zou et al., 2020), (Zou et al., 2018), and milder width assumptions (Ji & Telgarsky, 2019).

In contrast to the empirical success of local gradient methods, training a simple two-layer neural network with ReLU activation is proven to be NP-Hard (Boob et al., 2022). The sharp contrast between existing works on trainability guarantees and hardness results occurs as an increase in the width m, though it may seem to make the problem harder, actually makes it easier. Hence, it is a natural interest to understand the critical width m^* that guarantees polynomial time algorithms where $m \geq m^*$, and the corresponding algorithm.

Recently, it was shown that two-layer and deeper ReLU networks have exact convex reformulations (Pilanci & Ergen, 2020; Wang et al., 2022). However, the number of variables in these convex problems can be exponentially large. For a critical width $m^* \leq n + 1$, when $m \geq m^*$, we have an algorithm that exactly solves the non-convex problem with complexity polynomial in n (the number of data) and exponential in r (the rank of the dataset). Such an algorithm is given by considering the convex reformulation of the original problem which has approximately $(n/r)^r$ variables and solving the reformulated problem with standard interior-

[1]Department of Electrical Engineering, Stanford University, California, United States. Correspondence to: Sungyoon Kim <sykim777@stanford.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

point solvers. Due to the exponential factor in r, exactly solving the reformulated problem is intractable, and (Pilanci & Ergen, 2020) proposed to solve the randomly relaxed version of the exact reformulation with $m \ll (n/r)^r$ variables, where m is only polynomially large (see Section 2.1 for details). Though the relaxed problem has polynomial complexity with respect to all dimensions and works well in practice (see Figure 1), the optimality gap between the relaxed and the original problem has remained unknown. In Figure 1, width corresponds to the number of variables that the convex relaxation has.

In this paper, we study the randomized relaxation of the exact convex reformulation and find that:

- Under certain assumptions on the input data, the relative optimality gap between the non-convex problem and the randomized relaxation is bounded by a logarithmic factor. This shows that the convex relaxations run in polynomial-time and have strong approximation properties.
- Using results from (Wang et al., 2021) and (Li & Liang, 2018), we show that with high probability, local gradient methods with random initialization converges to a stationary point that has $O(\sqrt{\log n})$ relative training error with respect to the global minimum.
- We propose a tractable polynomial-time algorithm which is polynomial in all dimensions that can approximate the global optimum within a logarithmic factor. We are not aware of any similar result for regularized ReLU networks.

The paper is organized as follows: in Section 1.1, we go through a brief overview of related works. Section 2 is devoted to a detailed description of the contributions we made. Section 3 and Section 4 illustrate the overall proof strategy, novel approximation guarantees, and the geometry behind it: we first obtain approximation guarantees of an "easier version" of the randomly relaxed problem, and then solve the relaxed problem using the result. We wrap the paper with Section 5, presenting possible further discussions.

1.1. Prior and Related Work

Convex reformulation of neural networks: Starting from (Pilanci & Ergen, 2020), an extensive line of works have discussed the convex reformulation of neural networks. The main idea that lies in this line of work is that for different neural network architecture, e.g. for CNNs (Ergen & Pilanci, 2020), transformers (Sahiner et al., 2022), multi-layer networks (Ergen & Pilanci, 2021), vector outputs (Sahiner et al., 2020), etc, we have an exact convex reformulation with a different regularizer. Hence, neural network architectures can be understood as imposing different regularizations on the convex problem.

Moreover, these convex reformulations have given interesting insights into the training of neural networks. For example, (Wang et al., 2021) characterizes all stationary points of a two-layer neural network via global optimum of corresponding convex problems, and (Wang & Pilanci, 2023) discusses the intrinsic complexity of training two-layer networks with the view of convex duality and equivalence with the MAX-CUT problem.

Training complexity of two-layer neural networks: We use n to denote the number of data, d the dimension of data, and m to be the width of the model. Several papers have discussed the theoretical complexity of training simple ReLU networks and exact methods to train them. It is proven that for single-width neural networks, it is NP-Hard to train a two-layer ReLU network (Boob et al., 2022), and even approximating the optimal error with $(nd)^{1/poly(\log\log(nd))}$ is proven to be NP-Hard (Goel et al., 2020). (Dey et al., 2020) gives an approximate algorithm to solve the single-width neural network problem with $O(n^k)$ complexity and n/k approximate error for general input, and constant error bound in the student-teacher setting.

Some exact algorithms to train two-layer networks have been discussed, though due to the inherent complexity of training two-layer networks, these algorithms are often intractable in practice. (Arora et al., 2016) presents an algorithm with $O(2^m n^{dm} poly(n, d, m))$ complexity, where m is the width of the network. (Manurangsi & Reichman, 2018) shows that when both the input and weights are constrained on a unit ball, we can train a two-layer network to have ϵ - error bound with $O((2^{m/\epsilon})^{O(1)} n^{O(1)})$ complexity. Training guarantees in the student-teacher setting with Gaussian input have also been discussed: (Bakshi et al., 2019) discusses Gaussian inputs and in the student-teacher setting, we can exactly obtain the teacher parameters in $poly(n)$ time, and (Awasthi et al., 2021) discusses the same problem and propose an algorithm that works in $poly(n, m, d)$.

Note that the aforementioned complexity results are on training problems without regularization. When we have an l_2 regularized problem, the best exact solution to the problem is presented in (Pilanci & Ergen, 2020), where they propose an algorithm of $O(d^3(n/d)^{3d})$ complexity provided that the width of the network is sufficiently large. Extending the idea, (Bai et al., 2023) proposes an approximate method with complexity $O(d^2m^2)$ that works for $m \geq n/\xi$ for a predetermined error threshold ξ. However, their results require a very wide neural network and the relative optimal error bound is unclear. Note that choosing an appropriate regularization affects the performance of the model during test time and considering the regularized setting is not only for theoretical interest (see Appendix E for experiments).

Guarantees of local gradient methods for two-layer neural networks: An extensive line of work has discussed why local gradient methods such as gradient descent and its variants work so well for neural networks - even in the simplest setting of two layers. (Soudry & Carmon, 2016), (Soudry & Hoffer, 2017), (Tian, 2017) analyzes the local minima of neural networks with a similar flavor to our work: (Soudry & Carmon, 2016) shows when the number of parameters for a single layer exceeds n, differentiable local minima are global, and (Soudry & Hoffer, 2017) proves that differentiable regions that contain sub-optimal local minima are exponentially small compared to the ones containing global minima. (Tian, 2017) exploits the closed-form formula of population gradients to characterize the region of suboptimal critical points. Our work extends these works, for networks with regularization and non-differentiable stationary points under certain assumptions.

When Gaussian input is assumed, many works have analyzed whether local gradient methods can recover true parameters in the student-teacher setting. To prove that local gradient methods can recover true parameters, existing works either use a specifically tailored multi-phase analysis of gradient methods with deliberately chosen parameters and initialization (Li & Yuan, 2017), (Du et al., 2018), (Zhou et al., 2019), (Bao et al., 2024), or choose a particular structure of the model, e.g. no overlapping CNNs (Brutzkus & Globerson, 2017), (Zhang et al., 2020), two-layer network plus a skip connection (Li & Yuan, 2017), fixed second layer weights (Du et al., 2018), (Zhou et al., 2021). Our result is more abstract in the sense that it works for any local gradient method and initialization schemes that satisfy assumption (A2) and works in settings that are not student-teacher settings, though the analysis only gives relative approximation guarantees of these methods with respect to the global optimum.

Training guarantees have also been established for two-layer neural networks with different loss functions. Regarding hinge loss, (Brutzkus et al., 2017), (Laurent & Brecht, 2018), (Wang et al., 2019), (Wang & Pilanci, 2023) have analyzed the loss landscape of hinge loss under the assumption that the data is linearly separable. In particular, (Laurent & Brecht, 2018) discusses that all local minima are global for networks with leaky-ReLU activation, when we have hinge loss and the data is linearly separable, and (Wang et al., 2019) proposes a simple SGD-like algorithm that provably converges to a global minimum. (Wang & Pilanci, 2023) is worth noting that they have a similar convex analysis to ours to analyze the training hardness and approximation guarantees of training with hinge loss. Their guarantees are a different version of our result for hinge loss.

For useful lemmas that will be used throughout the paper, see Appendix A.

2. Main Results

2.1. Preliminaries

First, we go through a formal description of the optimization problems that we are interested in. Let the data matrix $X \in \mathbb{R}^{n\times d}$, the label vector $y \in \mathbb{R}^n$, and weight decay regularization $\beta > 0$. Consider the training problem

$$p^* := \min_{u_j,\alpha_j} \frac{1}{2}\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(\|u_j\|_2^2 + \|\alpha_j\|_2^2). \tag{1}$$

The convex counterpart of problem (1) can be written as in (Pilanci & Ergen, 2020):

$$p_0^* := \min_{u_i,v_i\in\mathcal{K}_{D_i}} \frac{1}{2}\|\sum_{D_i\in\mathcal{D}} D_iX(u_i - v_i) - y\|_2^2 + \beta\sum_{i\in\mathcal{I}}(\|u_i\|_2 + \|v_i\|_2), \tag{2}$$

where $\mathcal{D}$ is a set of hyperplane arrangement patterns $diag(\mathbb{1}[Xu \geq 0])$ and $\mathcal{K}_D = \{u \mid (2D - I)Xu \geq 0\}$ are cones where each optimization variable are constrained at. The intuition behind the convex counterpart is that as the nonconvexity of problem (1) comes from the nonlinearity and multiplying two variables u_j and α_j, we linearize the model and merge the two variables with appropriate scaling to make the problem convex (Mercklé et al., 2024).

Convex reformulation: Let's say $[n] = \{1, 2, ..., n\}$, D_i from $i \in [P]$ denote all possible hyperplane arrangement patterns, $\{u_i^*, v_i^*\}_{i=1}^P$ are global minima of problem (2),

$$m^* = \sum_{i=1}^{P} \mathbb{1}[u_i^* \neq 0] + \mathbb{1}[v_i^* \neq 0],$$

and $m \geq m^*$. Problems (1) and (2) then become equivalent, i.e. $p^* = p_0^*$ and we can construct an optimal solution of problem (1) with $\{u_i^*, v_i^*\}_{i=1}^P$. This means when $\mathcal{D} = \{D_i \mid i \in [P]\}$, we can exactly solve the original problem (1) with its convex reformulation (2). Though m^* might seem exponential in n as P is exponentially large in n, an application of Caratheodory's theorem shows that $m^* \leq n + 1$.

Gaussian relaxation: From (Cover, 1965), we know that $P = O((\frac{n}{r})^r)$ where $r = rank(X)$. Therefore, it is computationally intractable to sample all possible hyperplane arrangement patterns for large r. The randomized Gaussian relaxation of the problem (2) is where

$$\mathcal{D} = \{\tilde{D}_i = diag(\mathbb{1}[Xg_i \geq 0]) \mid g_i \sim \mathcal{N}(0, I_d), i \in [\tilde{P}]\},$$

i.e. instead of using every possible hyperplane arrangement pattern, we randomly sample $\tilde{P}$ patterns given by a random Gaussian vector. Note that we are using a very small portion

of the set of all hyperplane arrangement patterns, and it is not trivial that we will get good approximation guarantees.

Next, we illustrate an assumption on the data distribution for the main results.

Assumption on training data: Throughout the paper, we assume that the data distribution follows $X_{ij} \sim \mathcal{N}(0,1)$ i.i.d., $n/d = c$, where c is a constant, and denote this assumption as (A1). (A1) may be extended to a distribution with rotational symmetry, controllable quantile, and where Gordon's comparison (Thrampoulidis et al., 2014) or empirical process results (Mendelson et al., 2007) are applicable. A related analysis can be found in (Thrampoulidis & Hassibi, 2015), where they extend Gordon's comparison to isotropic random orthogonal matrices. Also, a connection to restricted isometry property (Candes, 2008) could be a key to extending the result to different distributions. Here, we assume Gaussianity for simplicity.
Also, the regime $n/d = c$, e.g. $n \asymp d$, is extensively studied in the literature (Montanari et al., 2019), (Celentano et al., 2021), (Celentano & Montanari, 2022). The given structure enables a quantitive comparison between the randomized relaxation and the original problem with random matrix theory and provides better bounds than arbitrary inputs.

2.2. Overview of Theoretical Results

We are now ready to state the main results of the paper. The first result is the optimality bound between the non-convex problem (1) and its Gaussian relaxation, assuming (A1).

Theorem 2.1. *(Informal) Consider the two-layer ReLU network training problem*

$$p^* := \min_{u_j, \alpha_j} \frac{1}{2}\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(\|u_j\|_2^2 + \|\alpha_j\|_2^2),$$

and its convex relaxation

$$\tilde{p}^* := \min_{u_i, v_i \in \mathcal{K}_{\tilde{D}_i}} \frac{1}{2}\|\sum_{\tilde{D}_i \in \mathcal{D}} D_i X(u_i - v_i) - y\|_2^2 + \beta\sum_{i \in \mathcal{I}}(\|u_i\|_2 + \|v_i\|_2). \quad (3)$$

Here, $|\mathcal{D}| = m/2$ and the elements are sampled by the hyperplane arrangement patterns of random Gaussian vectors. Assume (A1), d is sufficiently large, and suppose $m = \kappa \max\{m^*, 320(\sqrt{c}+1)^2 \log(\frac{n}{\delta})\}$ for some fixed $\kappa \geq 1$. Then, with high probability,

$$p^* \leq \tilde{p}^* \leq C\sqrt{\log 2n}\, p^*.$$

for some constant $C \geq 1$.

Remark 2.2. To the best of our knowledge, the above result provides the first polynomial-time approximation guarantee for regularized ReLU NNs. Also note that typically we have $p^* \to 0$ as $\beta \to 0$, implying $\tilde{p}^* - p^* \to 0$.

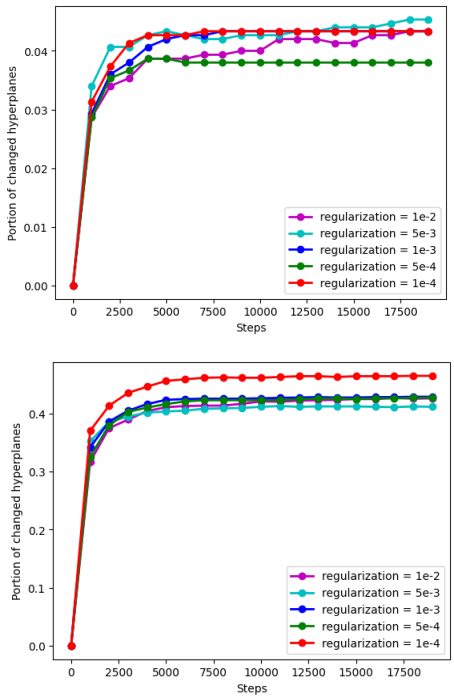

Figure 2. Verification of (A2) for gradient descent. The upper figure shows how many hyperplane arrangement patterns change for random data, and the lower figure shows how many of them change for MNIST.

For a detailed statement of the theorem, see Appendix D. A few points on Theorem 2.1 is worth mentioning. First, it gives a guarantee of the convex relaxation of the original problem when we sample only $\max\{m^*, O(\log n)\}$ hyperplane arrangement patterns. This is an exponential improvement over existing convex reformulations, and as m^* is much smaller than $n+1$ in practice, the width bound for Theorem 2.1 is practical. Also, with such characterization, we obtain a polynomial-time approximate algorithm that works with guarantees by simply solving the convex relaxation with standard interior-point solvers (Potra & Wright, 2000).

Theorem 2.3. *Assume (A1), d is sufficiently large, and suppose that $m = \kappa \max\{m^*, 320(\sqrt{c}+1)^2 \log(n/\delta)\}$ for some fixed $\kappa \geq 1$. Then, there exists a randomized algorithm with $O(d^3m^3)$ complexity that solves problem (1) within $O(\sqrt{\log n})$ relative optimality bound with high probability.*

Moreover, with the characterization of stationary points in (Wang et al., 2021), we can prove that local gradient methods converge to "nice" stationary points under the assumption below:

(A2) While local gradient method iterates, only a randomly chosen portion $p < 1/2$ of the initial hyperplane arrangement patterns change until it converges. (A2) is also con-

sidered in (Li & Liang, 2018), where they show that SGD with random initialization preserves most of the hyperplane arrangement patterns. The specific lemma, which they refer to as the coupling lemma, is proven for cross-entropy loss. It is also verified for gradient descent with MSE loss, both for random data and MNIST in Figure 2.

Lemma 2.4. *(Coupling lemma, (Li & Liang, 2018)) With high probability over the random initialization that follows $\mathcal{N}(0, m^{-1}I)$, and suppose the data is generated from l underlying distributions. For every $\tau > 0$, $t = \tilde{O}(\tau/\eta)$, at least $1 - e\tau l/\sigma$ portion of the hyperplane arrangement patterns remain the same.*

The idea is that we can map each stationary point to a global minimum of a convex problem, and as we initialize at random, we can think of the stationary points as global minimizers of the convex problem with randomized hyperplane arrangement patterns.

Theorem 2.5. *(Informal) Consider the training problem $\min_{u_j,\alpha_j} \mathcal{L}(u, \alpha)$, where the loss function $\mathcal{L}$ is given as*

$$\mathcal{L}(u, \alpha) = \frac{1}{2}\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(\|u_j\|_2^2 + \alpha_j^2).$$

Assume (A1),(A2) and d, m are sufficiently large. For any random initialization $\{u_i^0, \alpha_i^0\}_{i=1}^m$, suppose local gradient method converged to a stationary point $\{u_i', \alpha_i'\}_{i=1}^m$. Then, with high probability,

$$\mathcal{L}(u', \alpha') \leq C\sqrt{\log 2n}\, \mathcal{L}(u^*, \alpha^*),$$

for some $C \geq 1$. Here, $\{u_i^*, \alpha_i^*\}_{i=1}^m$ is a global optimum of $\mathcal{L}(u, \alpha)$.

Remark 2.6. The above result shows that there exists many stationary points whose objective value is a logarithmic approximation of the global optimum. Therefore, first-order optimizers such as SGD and Adam can approximate the global optimum even when they converge to stationary points.

Note that our analysis is not based on explicit iteration of local gradient methods, or does not exploit a NTK - based analysis. Rather, it follows from a simple yet clear analysis based on the lens of convex optimization.

2.3. Notations for Proof

Before discussing the proof strategy, we clarify some frequently used variables in this section. We use D_i for $i \in [P]$ to denote all possible hyperplane arrangements, P to denote the number of all possible hyperplane arrangements, $\tilde{D}_i$ to denote randomly selected hyperplane arrangement patterns, and $\tilde{P}$ the number of such subsamples. We also use $\mathcal{K}_D$ as in preliminaries. $\mathcal{M}$ is used to denote

$$\mathbb{E}_{g\sim\mathcal{N}(0,I_d)}[diag[\mathbb{1}(Xg \geq 0)]XX^T diag[\mathbb{1}(Xg \geq 0)]],$$

and κ, which is analogous to the condition number, to denote

$$\frac{\lambda_{max}(XX^T)}{\lambda_{min}(\mathcal{M})},$$

provided that $\mathcal{M}$ is invertible. At last, m^* is used to denote the number of nonzero variables for the optimal solution of the convex reformulation.

2.4. Overall Proof Strategy

To prove Theorem 2.1, we first consider the unconstrained problem

$$\tilde{p_1}^* := \min_{u_i, v_i \in \mathbb{R}^d} \frac{1}{2}\|\sum_{\tilde{D}_i \in \mathcal{D}} \tilde{D}_i X(u_i - v_i) - y\|_2^2 + \beta\sum_{i\in\mathcal{I}}(\|u_i\|_2 + \|v_i\|_2). \quad (4)$$

The convex problem (4) was first introduced in (Mishkin et al., 2022b), where they first solve the unconstrained problem and decompose it into cone constraints to solve the original problem (3). The problem is further equivalent to

$$\min_{u_i\in\mathbb{R}^d} \frac{1}{2}\|\sum_{\tilde{D}_i\in\mathcal{D}} \tilde{D}_i X w_i - y\|_2^2 + \beta\sum_{i\in\mathcal{I}}\|w_i\|_2, \quad (5)$$

due to triangle inequality. We use a similar strategy that was introduced in (Mishkin et al., 2022b): first, we show that the unconstrained problem enjoys good approximation guarantees, even though we use only a subsample of the whole possible hyperplane arrangement patterns. After that, for global optimum $\{w_i^*\}_{i=1}^m$ of problem (5), we find $u_i, v_i \in \mathcal{K}_{D_i}$ that satisfies $u_i - v_i = w_i^*$ and minimal norm sum $\|u_i\|_2 + \|v_i\|_2$. We find that this decomposition does not increase the norm up to $O(\sqrt{\log 2n})$ factor, finally leading to the guarantee of the original problem. A novel duality-based analysis of the unconstrained problem and obtaining a relative optimality gap (Corollary 3.10) and an analysis using Gordon's comparison to understand how "sharp" each cone may be (Corollary 4.5) are the major technical contributions of our paper.

3. Guarantees for the Unconstrained Convex Relaxation

Let's start with the unconstrained optimization problem Equation (5). Note $\tilde{p}_1^*$ to be the optimal value of the random relaxation of the unconstrained problem, and p_1^* to be the optimal value of the unconstrained problem using all possible hyperplane arrangement patterns. We wish to find constants $C_1, C_2 > 0$ depending only with the data matrix X that satisfies

$$0 \leq \tilde{p}_1^* - p_1^* \leq C_1, \quad p_1^* \leq \tilde{p}_1^* \leq C_2 p_1^*.$$

As a warmup, we first give results when the problem is unregularized. We can show that $p_1^* = \tilde{p}_1^* = 0$ with high probability when we sample sufficiently many hyperplane arrangement patterns. After that, we give approximation results for a gated ReLU problem with l_2 regularization and use them to prove that

$$C_1 = \sqrt{2}\beta\frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}}, C_2 = 2\sqrt{2}\sqrt{\frac{\lambda_{\max}(XX^T)}{\lambda_{\min}(\mathcal{M})}},$$

holds with probability at least $1-\delta$, provided that we sample sufficiently many hyperplanes and the dimension d is sufficiently large. We defer the proofs to Appendix B.

3.1. Warmup: Unconstrained Relaxation Without Regularization

To show that $p_1^* = 0$, we prove that we can approximate any vector with $2n$ hyperplane arrangement patterns. The proof is consistent with general overparametrization arguments, which state that when the width of the network is $m \geq n$, any local minimum becomes the global minimum. Here, we carefully choose $2n$ hyperplane arrangement patterns where each arrangement pair $D_{1,i}, D_{2,i}$ differ only at the i-th diagonal entry. With n such pairs we can express any vector $y \in \mathbb{R}^n$ as a linear combination of vectors in the column space of $\{(D_{1,i} - D_{2,i})u | u \in \mathbb{R}^n\} = \{ke_i | k \in \mathbb{R}\}$, proving that there exists u_i s satisfying

$$\sum_{i=1}^{n}(D_{1,i} - D_{2,i})Xu_i = y.$$

This means that if we have all possible hyperplane arrangement patterns we can express any vector y with zero error. The specific proposition is as follows.

Proposition 3.1. *Suppose no two rows of X are parallel. There exists hyperplane patterns $\tilde{D}_1$, $\tilde{D}_2$, ..., $\tilde{D}_{2n}$ such that for any $y \in \mathbb{R}^n$, there exists $u_1, u_2, ...u_{2n} \in \mathbb{R}^d$ satisfying*

$$\sum_{i=1}^{2n}\tilde{D}_i X u_i = y.$$

However, we can further prove that using randomly sampled hyperplane arrangement patterns $\tilde{D}_i$, we can fit arbitrary vector y with probability at least $1-\delta$, provided that we sample sufficiently many hyperplane arrangement patterns.

Proposition 3.2. *Suppose we sampled $\tilde{P} = 2\kappa\log(\frac{n}{\delta})$ hyperplane arrangement patterns, provided that $\mathcal{M}$ is invertible. Then, with probability at least $1-\delta$, for any $y \in \mathbb{R}^n$ there exists $u_1, u_2, ..., u_{\tilde{P}}$ satisfying*

$$\sum_{i=1}^{\tilde{P}}\tilde{D}_i X u_i = y.$$

Proposition 3.2 shows that when we sample sufficiently many hyperplane arrangement patterns, we can fit arbitrary vectors with high probability, hence $\tilde{p}_1^* = 0$. The two propositions lead to the following approximation result.

Corollary 3.3. *When $\beta = 0$ and $\tilde{P} \geq 2\kappa\log(\frac{n}{\delta})$, $p_0^* = p_1^* = 0$ with probability at least $1-\delta$.*

One natural question is "how large will κ be?" Interestingly, when (A1) is satisfied, we can show that $\kappa = O(n/d)$ holds (see Section 3.5 for details). Hence, for random data, we use approximately $O(n\log n)$ parameters to fit to $y \in \mathbb{R}^n$ with high probability. Note that we need at least n/d hyperplane arrangement patterns and n parameters to fit to any vector. This means that for random data, Proposition 3.2 is optimal up to a logarithmic factor.

3.2. Unconstrained Relaxation with l_2 Regularization

In the case of l_2 regularization, we can find that when we sample more hyperplane arrangement patterns, the optimal value decreases with the same order as the number of planes with high probability. The proof is rather straightforward: we solve the linear-constrained quadratic problem, and then use matrix chernoff bounds to relate the eigenvalues of $\mathcal{M}$ with the eigenvalue of sample mean of matrices

$$\mathcal{M}_{\tilde{P}} = \frac{1}{\tilde{P}}\sum_{i=1}^{\tilde{P}}\tilde{D}_i XX^T\tilde{D}_i.$$

Theorem 3.4. *Let p_2^* be the optimal value of the unconstrained relaxed problem with l_2 regularization, and suppose $\tilde{P} \geq 12\kappa\log(\frac{2n}{\delta})$. With probability at least $1-\delta$ there exists scalars $C_1, C_2 > 0$ that satisfies*

$$\frac{C_1}{\tilde{P}} \leq p_2^* \leq \frac{C_2}{\tilde{P}}.$$

An intuitive explanation of Theorem 3.4 is that when we sample sufficiently many hyperplane arrangement patterns, we have $\mathcal{M}_{\tilde{P}} \approx \mathcal{M}$ and the optimal value of the l_2 regularized problem concentrates to $\frac{\beta}{\tilde{P}}y^T\mathcal{M}^{-1}y$. This characterization means that for l_2 regularization, we cannot have a relative optimality gap that is independent of P, the total number of all possible hyperplane arrangement patterns.

3.3. Unconstrained Relaxation With Group l_1 Regularization

Different from the case with l_2 regularization, when we have group l_1 regularization, we can have relative error bounds that are irrelevant to the total number of hyperplane arrangement patterns. We first obtain an upper bound of $\tilde{p}_1^*$ by using surrogate variables κ_i and solving the l_2 regularized

problem,

$$\min_{\kappa_i \in \mathbb{R}} \min_{u_i \in \mathbb{R}^d} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y \|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} (\kappa_i \|u_i\|_2^2 + \frac{1}{\kappa_i}). \tag{6}$$

We cannot exactly solve problem (6). However, we can use the ansatz $\kappa_i = \frac{\kappa}{\tilde{P}}$ and apply matrix concentration to upper bound the optimal value.

Proposition 3.5. *Let $\tilde{p}_1^*$ the optimal value of the unconstrained relaxation Equation (5). Suppose we sampled $\tilde{P} \geq 8\kappa \log(\frac{n}{\delta})$ hyperplane arrangement patterns. With probability at least $1-\delta$,*

$$\tilde{p}_1^* \leq \sqrt{2}\beta \frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}},$$

provided that $\mathcal{M}$ is invertible.

Then, we lower bound p_1^* using the dual problem of Equation (5). As the dual problem becomes a maximization of the dual objective with respect to a dual variable λ, choosing λ well can give a meaningful lower bound on p_1^*. In the proof we choose a scalar multiple of y to obtain a bound that has a similar scale with $\sqrt{2}\beta \frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}}$.

Proposition 3.6. *Let p_1^* be the optimal value of the unconstrained problem using all hyperplane arrangement patterns. When we write $\mathcal{M}_i = D_i X X^T D_i$,*

$$G\beta \frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}} \leq p_1^*,$$

where

$$G = 1 - \frac{\beta}{2 \max_{i \in [P]} \sqrt{y^T \mathcal{M}_i y}}.$$

A direct corollary simplifies the lower bound. Corollary 3.7 can be used when β is sufficiently smaller than $\|y\| \sqrt{\lambda_{\min}(\mathcal{M})}$. As $\|y\|$ is the scale of $O(\sqrt{n})$ and $\lambda_{\min}(\mathcal{M})$ grows with d for random data matrix X, it is likely that the value $\|y\|_2 \sqrt{\lambda_{\min}(\mathcal{M})}$ is much larger than β for most cases.

Corollary 3.7. *Suppose further that $\|y\|_2 \sqrt{\lambda_{\min}(\mathcal{M})} \geq \beta$. Then,*

$$\frac{\beta}{2} \frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}} \leq p_1^*.$$

By using Proposition 3.5 and Proposition 3.6, we obtain a relative error bound between p_1^* and $\tilde{p}_1^*$.

Theorem 3.8. *Let p_1^* and $\tilde{p}_1^*$ be optimal values of problem (5) with all possible hyperplane arrangements and randomly sampled arrangements, respectively. Suppose we sampled $\tilde{P} \geq 8\kappa \log(\frac{n}{\delta})$ hyperplane arrangement patterns and $\mathcal{M}$ is invertible. We have*

$$0 \leq \tilde{p}_1^* - p_1^* \leq \sqrt{2}\beta \frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}},$$

and

$$p_1^* \leq \tilde{p}_1^* \leq \frac{\sqrt{2\kappa}}{G} p_1^*,$$

with probability at least $1-\delta$.

Note that different from the l_2 regularized case, in this case, we have a relative error bound that is independent of the total number of hyperplane arrangement patterns.

A better approximation is possible if we take into account two things: one is that we can use $\max_i \sqrt{y^T \mathcal{M}_i y}/\|y\|_2$ instead of $\lambda_{\max}(XX^T)$ to bound the relative error with a tighter bound, and we can also use $\mathcal{M}_{\tilde{P}}$ directly instead of $\lambda_{\min}(\mathcal{M})$. Hence, the overall bound can be improved to

$$p_0^* \leq p_1^* \leq \frac{\sqrt{2}}{G} \frac{\max_i \sqrt{y^T \mathcal{M}_i y} \sqrt{y^T \mathcal{M}_{\tilde{P}}^{-1} y}}{\|y\|_2^2} p_0^*, \tag{7}$$

where we sampled sufficiently many hyperplane arrangement patterns to make $\mathcal{M}_{\tilde{P}}$ invertible.

3.4. Connection to the MAX-CUT Problem

There is an interesting connection between the upper bound of the relative error and the MAX-CUT problem. From bound (7), we can bound

$$\max_i \sqrt{y^T \mathcal{M}_i y} \leq \sqrt{\max_{b \in \{0,1\}^n} b^T diag(y) X X^T diag(y) b},$$

where solving the right-hand side is equivalent to solving the max-cut problem of the graph with an adjacency matrix

$$\frac{1}{4} \begin{bmatrix} I \\ 1^T \end{bmatrix} diag(y) X X^T diag(y) \begin{bmatrix} I & 1 \end{bmatrix}.$$

Hence, in special cases where X is orthogonally separable or when X has negative correlation, we may further bound the relative error. The connection with MAX-CUT shows that the relative error, or the possibility of approximation, is intrinsically related to the structure of the dataset and its clusterization properties.

3.5. Scale of κ for Random Data

In this subsection, we show that under (A1), we can show that $\kappa \leq O(c)$ holds with high probability for sufficiently large d. This means that for random data, as long as d grows with n, we only need to sample $O(\log n)$ hyperplane arrangement patterns to obtain a constant factor approximation of the unconstrained relaxation. This bound is more practical compared to most bounds that require networks of width at least n.

To prove the identity, we first use the non-asymptotic inequality deduced from Gordon's minimax comparison

$$\lambda_{max}(XX^T) \leq 2(\sqrt{n} + \sqrt{d})^2,$$

which holds with high probability. To upper bound κ, finding a lower bound of $\lambda_{\min}(\mathcal{M})$ is enough.

To lower bound $\lambda_{\min}(\mathcal{M})$, we find a closed form expression of $\mathcal{M}_{ij}$ and use Lemma A.4 to approximate $\mathcal{M}$ via linearization. We need d to be sufficiently large to apply concentration inequalities on $\|x_i\|_2$, the norm of rows of X.

Theorem 3.9. *Assume (A1) and suppose $c \geq 1, \delta > 0$. There exists d_1 such that if $d \geq d_1$, with probability at least $1 - \delta'$, we have*

$$\lambda_{min}(\mathcal{M}) \geq \frac{d}{10}.$$

Hence, κ is indeed upper bounded with a constant factor when d is sufficiently large. Using the fact directly leads to the following corollary.

Corollary 3.10. *When the conditions of Theorem 3.9 holds, with probability at least $1 - \delta' - e^{-Cn}$, we have*

$$\kappa \leq 20(\sqrt{c}+1)^2,$$

for some $C > 0$. Moreover, let p_1^* and $\tilde{p}_1^*$ be optimal values of problem (5) with all possible hyperplane arrangements and randomly sampled arrangements, respectively. When we sample $\tilde{P} \geq 160(\sqrt{c}+1)^2 \log(\frac{n}{\delta})$ hyperplane arrangement patterns, we have

$$p_1^* \leq \tilde{p}_1^* \leq \frac{2\sqrt{10}}{G}(\sqrt{c}+1)p_1^*$$

with probability at least $1 - \delta - \delta' - e^{-Cn}$.

4. Extension to the Constrained Problem

In this section, we move a step further by discussing the relative error between the optimal value of the convex reformulation and its random relaxation. Let $\{w_i^*\}_{i=1}^m$ be the solution of the unconstrained relaxation of the original problem and let $u_i^*, v_i^* \in \mathcal{K}_{D_i}$ satisfying $w_i^* = u_i^* - v_i^*$ with minimal $\|u_i^*\|_2 + \|v_i^*\|_2$. Although there is no universal constant C that satisfies

$$\|u_i^*\|_2 + \|v_i^*\|_2 \leq C\|w_i^*\|_2,$$

for any cone $\mathcal{K}$ (Mishkin et al., 2022b), we can show that for random data X, it is likely that there exists reasonably large $\mathcal{C}$ that satisfies

$$\|u_i^*\|_2 + \|v_i^*\|_2 \leq \mathcal{C}\|w_i^*\|_2,$$

for randomly sampled cones $\mathcal{K}_{\tilde{D}_i}$ and all w_i^* with high probability. In 2 dimensions, $\mathcal{C}$ is directly related to the angle between two rays of the convex cone. Hence, $\mathcal{C}$ is analogous to the 'sharpness of each cone $\mathcal{K}_{\tilde{D}_i}$'. We defer the proofs to Appendix C.

4.1. Cone Sharpness $\mathcal{C}$

We start by defining the cone sharpness constant $\mathcal{C}(\mathcal{K}, z)$ for a given convex cone $\mathcal{K}$.

Definition 4.1. For a cone $\mathcal{K}$ and unit vector z, the sharpness with respect to z is defined as

$$C(\mathcal{K}, z) = \min_{u,v \in \mathcal{K}, u-v=z} \|u\|_2 + \|v\|_2.$$

We can upper bound the unconstrained relaxation with the constrained relaxation using the cone sharpness.

Proposition 4.2. *Let $\tilde{p}_0^*$ and $\tilde{p}_1^*$ be optimal values of the Gaussian relaxation of the convex reformulation and its unconstrained version, respectively. Suppose the unconstrained problem has solutions w_i^* for $i \in [\tilde{P}]$, and let*

$$C = \max_{i \in [\tilde{P}]} \mathcal{C}(\mathcal{K}_{\tilde{D}_i}, \frac{w_i^*}{\|w_i^*\|_2}).$$

Then, $\tilde{p}_0^* \leq C\tilde{p}_1^*$ holds.

Now, we introduce a strategy to upper bound $\mathcal{C}(\mathcal{K}_{\tilde{D}_i}, z)$. An interesting fact is that when (A1) is satisfied, the cone sharpness constant is not large with high probability, and can be bounded by $O(\sqrt{\log n})$ factor. The intuition here is that when $n \asymp d$, we have approximately $O((n/d)^d)$ many cones - which is only exponential in d when $n/d = c$. Hence, in d dimensions, the number of cone constraints is not that many, and we can have reasonable upper bounds on the sharpness.

We first start with a simple proposition that upper bounds $C(\mathcal{K}_{\tilde{D}_i}, z)$. The idea is similar to Chebyshev centers of a polyhedron (Boyd & Vandenberghe, 2004).

Proposition 4.3. *Take any unit vector z, and suppose $\mathcal{K} = \{u | (2D - I)Xu \geq 0\}$. If there exists vector u that satisfies*

$$\|u\|_2 \leq 1, \quad (2D - I)Xu \geq \epsilon \cdot |(2D - I)Xz|,$$

we know that

$$C(\mathcal{K}, z) \leq 1 + \frac{1}{\epsilon}.$$

When (A1) is satisfied and for a randomly sampled cone $\mathcal{K}_{\tilde{D}_i}$, with the rotational invariance of X, we can rotate the cone to contain e_1 and the distribution of other elements on X will not change. Therefore, without loss of generality, we may assume that $\mathcal{K} = \{u \mid \tilde{X}u \geq 0\}$, where the first column of $\tilde{X}$ is the absolute value sampled from $\mathcal{N}(0, 1)$ and the other columns are sampled from $\mathcal{N}(0, 1)$.

We wish to construct a vector u that satisfies the conditions in Proposition 4.3. Using a novel application of Gordon's comparison, we find such a vector for any unit vector z.

Theorem 4.4. *Let $b \in \mathbb{R}^n$ sampled from the folded normal distribution, and $X \in \mathbb{R}^{n\times d}$ be a matrix where each entries are sampled from a normal distribution. Consider the random variable*

$$F(X,b) = \max_{\|z\|_2=1} \min_{\substack{Xu\geq -Xz-kb \\ Xu\geq Xz-kb \\ k\geq 0}} \|u\|_2 + k,$$

where $u, z \in \mathbb{R}^d$. Then, with probability at least $1-1/n^{10}-e^{-Cd}$ for some positive constant C,

$$F(X,b) \leq 200c\sqrt{c\log 2n}. \tag{8}$$

For any given unit vector z, we can construct u for the rotated cone $\mathcal{K} = \{u \mid \tilde{X}u \geq 0\}$ by solving the minimization problem in Theorem 4.4. Eventually, we can bound $\mathcal{C}(\mathcal{K}, z)$ with a logarithmic factor.

Corollary 4.5. *Suppose n, d are sufficiently large that Equation (8) holds with probability at least $1-\delta''$, for $b \in \mathbb{R}^n$ sampled from a folded normal distribution and $X \in \mathbb{R}^{n\times d-1}$ sampled from a normal distribution. Then, with probability at least $1-\delta''$,*

$$C(\mathcal{K}_{\tilde{D}_i}, z) \leq 2 + 200c\sqrt{c\log 2n},$$

also holds for all unit vectors z.

A direct corollary is that we can approximate the convex reformulation with its unconstrained version, having $O(\sqrt{\log n})$ scale relative bound (Corollary C.5).

4.2. Proof of the Main Results

After the bound on the cone sharpness, the proof of the main theorems follows almost immediately. By considering the convex reformulation of the original problem (1), and first upper bounding it with the unconstrained problem, then upper bounding it again with the unconstrained problem with random relaxation, we can find a relative optimality gap between the original problem and its convex relaxation. It is clear that we immediately get a tractable polynomial-time randomized algorithm by solving the randomly relaxed convex problem. At last, by identifying stationary points of the original problem as the global minimum of randomly subsampled convex problems, we obtain Theorem 2.5. See Appendix D for a detailed proof.

5. Conclusion

In this paper, we provided guarantees of approximating the equivalent convex program given in (Pilanci & Ergen, 2020) with a much smaller random subprogram. With assumptions on X and the dimension d, we proved that the optimal value of the subsampled convex program approximates the full convex program up to a logarithmic factor. Using the approximation results we discuss novel insights on training two-layer neural networks, by showing that under mild assumptions local gradient methods converge to stationary points with optimality guarantees, and propose a practical algorithm to train neural networks with guarantees.

We hope to improve the work in two ways: First, removing the logarithmic factor of the approximation would be an important problem to tackle. Also, extending the theorems to different architectures, i.e. CNNs, transformers, and multi-layer networks would be meaningful.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant DMS-2134248; in part by the NSF CAREER Award under Grant CCF-2236829; in part by the U.S. Army Research Office Early Career Award under Grant W911NF-21-1-0242; and in part by the Office of Naval Research under Grant N00014-24-1-2164.

Impact statement

Our work advances the understanding training of neural networks, and we feel no societal consequences should be specifically highlighted within the paper.

References

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Awasthi, P., Tang, A., and Vijayaraghavan, A. Efficient algorithms for learning depth-2 neural networks with general relu activations. *Advances in Neural Information Processing Systems*, 34:13485–13496, 2021.

Bai, Y., Gautam, T., and Sojoudi, S. Efficient global optimization of two-layer relu networks: Quadratic-time algorithms and adversarial training. *SIAM Journal on Mathematics of Data Science*, 5(2):446–474, 2023.

Bakshi, A., Jayaram, R., and Woodruff, D. P. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pp. 195–268. PMLR, 2019.

Bao, Y., Shehu, A., and Liu, M. Global convergence analysis of local sgd for two-layer neural network without

overparameterization. *Advances in Neural Information Processing Systems*, 36, 2024.

Boob, D., Dey, S. S., and Lan, G. Complexity of training relu neural network. *Discrete Optimization*, 44:100620, 2022.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brutzkus, A. and Globerson, A. Globally optimal gradient descent for a convnet with gaussian inputs. In *International conference on machine learning*, pp. 605–614. PMLR, 2017.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Candes, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10):589–592, 2008.

Celentano, M. and Montanari, A. Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics*, 50(1):170–196, 2022.

Celentano, M., Cheng, C., and Montanari, A. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965. doi: 10.1109/PGEC.1965.264137.

Dey, S. S., Wang, G., and Xie, Y. Approximation algorithms for training one-node relu neural networks. *IEEE Transactions on Signal Processing*, 68:6696–6706, 2020.

Du, S., Lee, J., Tian, Y., Singh, A., and Poczos, B. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pp. 1339–1348. PMLR, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

El Karoui, N. The spectrum of kernel random matrices. 2010.

Ergen, T. and Pilanci, M. Implicit convex regularizers of cnn architectures: Convex optimization of two-and three-layer networks in polynomial time. *arXiv preprint arXiv:2006.14798*, 2020.

Ergen, T. and Pilanci, M. Global optimality beyond two layers: Training deep relu networks via convex programs. In *International Conference on Machine Learning*, pp. 2993–3003. PMLR, 2021.

Goel, S., Klivans, A., Manurangsi, P., and Reichman, D. Tight hardness results for training depth-2 relu networks. *arXiv preprint arXiv:2011.13550*, 2020.

Harvey, N. Lecture 2: Matrix chernoff bounds.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Laurent, T. and Brecht, J. The multilinear structure of relu networks. In *International conference on machine learning*, pp. 2908–2916. PMLR, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 30, 2017.

Manurangsi, P. and Reichman, D. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

Mercklé, V., Iutzeler, F., and Redko, I. The hidden convex optimization landscape of two-layer relu networks. In *ICLR Blogposts 2024*, 2024. URL https://iclr-blogposts.github.io/2024/blog/hidden-convex-relu/. https://iclr-blogposts.github.io/2024/blog/hidden-convex-relu/.

Mishkin, A., Sahiner, A., and Pilanci, M. Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions. *CoRR*, abs/2202.01331, 2022a. URL https://arxiv.org/abs/2202.01331.

Mishkin, A., Sahiner, A., and Pilanci, M. Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions. In *International Conference on Machine Learning*, pp. 15770–15816. PMLR, 2022b.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Pilanci, M. and Ergen, T. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pp. 7695–7705. PMLR, 2020.

Potra, F. A. and Wright, S. J. Interior-point methods. *Journal of computational and applied mathematics*, 124(1-2): 281–302, 2000.

Sahiner, A., Ergen, T., Pauly, J., and Pilanci, M. Vector-output relu neural network problems are copositive programs: Convex analysis of two layer networks and polynomial-time algorithms. *arXiv preprint arXiv:2012.13329*, 2020.

Sahiner, A., Ergen, T., Ozturkler, B., Pauly, J., Mardani, M., and Pilanci, M. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. In *International Conference on Machine Learning*, pp. 19050–19088. PMLR, 2022.

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Soudry, D. and Hoffer, E. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

Thrampoulidis, C. and Hassibi, B. Isotropically random orthogonal matrices: Performance of lasso and minimum conic singular values. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 556–560. IEEE, 2015.

Thrampoulidis, C., Oymak, S., and Hassibi, B. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Tian, Y. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International conference on machine learning*, pp. 3404–3413. PMLR, 2017.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Wang, G., Giannakis, G. B., and Chen, J. Learning relu networks on linearly separable data: Algorithm, optimality, and generalization. *IEEE Transactions on Signal Processing*, 67(9):2357–2370, 2019.

Wang, Y. and Pilanci, M. Polynomial-time solutions for relu network training: A complexity classification via max-cut and zonotopes. *arXiv preprint arXiv:2311.10972*, 2023.

Wang, Y., Lacotte, J., and Pilanci, M. The hidden convex optimization landscape of regularized two-layer relu networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2021.

Wang, Y., Ergen, T., and Pilanci, M. Parallel deep neural networks have zero duality gap. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhang, S., Wang, M., Xiong, J., Liu, S., and Chen, P.-Y. Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2622–2635, 2020.

Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E., and Zhao, T. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pp. 7594–7602. PMLR, 2019.

Zhou, M., Ge, R., and Jin, C. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pp. 4577–4632. PMLR, 2021.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

A. Useful Lemmas

Lemma A.1. *(Matrix Chernoff (Tropp et al., 2015), (Harvey)) Let $X_1, X_2, ..., X_k$ be independent, random, symmetric real $n \times n$ matrix with $0 \preccurlyeq X_i \preccurlyeq RI$. Let $\mu_{min} I \preccurlyeq \sum_{i=1}^{k} \mathbb{E}[X_i] \preccurlyeq \mu_{max} I$. Then, for all $\delta \in [0, 1]$,*

$$\mathbb{P}(\lambda_{max}(\sum_{i=1}^{k} X_i) \geq (1+\delta)\mu_{max}) \leq n e^{-\delta^2 \mu_{max}/3R}$$

and

$$\mathbb{P}(\lambda_{min}(\sum_{i=1}^{k} X_i) \leq (1-\delta)\mu_{min}) \leq n e^{-\delta^2 \mu_{min}/2R}$$

holds.

Lemma A.2. *(Gordon's comparison (Thrampoulidis et al., 2014)) Let $\Phi(X)$ be*

$$\Phi(X) = \min_{x \in S_x} \max_{y \in S_y} y^T X x + \psi(x, y),$$

where X is a random matrix with i.i.d. Gaussian entries, and $\phi(g, h)$ be

$$\phi(g, h) = \min_{x \in S_x} \max_{y \in S_y} \|x\|_2 g^T y + \|y\|_2 h^T x + \psi(x, y),$$

where entries of g, h are i.i.d. and sampled from $\mathcal{N}(0, 1)$. Then, we have

$$\mathbb{P}(\Phi(X) \leq M) \leq 2\mathbb{P}(\phi(g, h) \leq M).$$

Lemma A.3. *(Weyl's theorem on eigenvalues) Suppose $A, B \in \mathbb{R}^{n \times n}$ are symmetric. Then, the below estimates on the minimum eigenvalue of A, B hold.*

$$\lambda_{min}(A) \geq \lambda_{min}(B) - \|A - B\|_F, \tag{9}$$

$$\lambda_{min}(A + B) \geq \lambda_{min}(A) + \lambda_{min}(B). \tag{10}$$

Here, $\rho(A)$ denotes the spectral radius of A, and $\|A\|_F$ is the Frobenius norm of A.

Lemma A.4. *(Spectrum of kernel random matrices (El Karoui, 2010)) Suppose we obtain n i.i.d. vectors x_i from $\mathcal{N}(0, I_d)$. Let's consider the kernel matrix*

$$K_{i,j} = f(\frac{x_i \cdot x_j}{d}).$$

We assume that:
(a) $n/d, d/n$ are bounded as $d \to \infty$.
(b) f is a C^1 function in a neighborhood of 1, and a C^3 function in a neighborhood of 0.
Under the assumptions, the kernel matrix M can (in probability) be approximated consistently in operator norm, when $d, n \to \infty$, by the matrix K', where

$$K' = (f(0) + \frac{f''(0)}{2d}) 1 1^T + \frac{f'(0)}{d} X X^T + (f(1) - f(0) - f'(0)) I_n.$$

B. Proofs in Section 3.

Proposition B.1. *(Proposition 3.1. of the paper) Suppose no two rows of X are parallel. There exists hyperplane patterns $\tilde{D}_1$, $\tilde{D}_2$, ..., $\tilde{D}_{2n}$ such that for all $y \in \mathbb{R}^n$, there exists $u_1, u_2, ...u_{2n} \in \mathbb{R}^d$ satisfying*

$$\sum_{i=1}^{2n} \tilde{D}_i X u_i = y.$$

Proof. Let's begin with a simple observation: for any $i \in [n]$, there exists two different hyperplane pattern $D_{1,i}$ and $D_{2,i}$ satisfying

$$|D_{1,i} - D_{2,i}| = diag(e_i).$$

The construction is intuitive when thought geometrically, passing through only the ith plane. A rigorous construction is as below: First, choose a point $c \in \mathbb{R}^d$ satisfying

$$c \cdot X_i = 0, \quad c \cdot X_j \neq 0 \quad for\ all \quad j \neq i,$$

where $X_1, X_2, ..., X_n$ are rows of X. We can choose such point because the set $\{u|u \cdot X_i = 0\} \cap \{u|u \cdot X_j = 0\}$ has measure 0 in $\{u|u \cdot X_i = 0\}$, provided no two rows of X are parellel. Now, we choose ϵ to be small such that

$$\epsilon = \min_{j \neq i, X_j \cdot X_i \neq 0} \frac{|c \cdot X_j|}{2|X_i \cdot X_j|} > 0.$$

and take two points $c + \epsilon X_i$, $c - \epsilon X_i$. Now, for the ith hyperplane,

$$(c + \epsilon X_i) \cdot X_i > 0, \quad (c - \epsilon X_i) \cdot X_i < 0,$$

and for all other hyperplanes $j \neq i$,

$$\epsilon |X_i \cdot X_j| \leq \frac{1}{2} |c \cdot X_j|,$$

thus the sign of $(c + \epsilon X_i) \cdot X_j, c \cdot X_j, (c - \epsilon X_i) \cdot X_j$ are identical. This means the hyperplane patterns of $c + \epsilon X_i$ and $c - \epsilon X_i$ differ only for the ith plane, meaning we have found two planes $D_{1,i}, D_{2,i}$ such that $|D_{1,i} - D_{2,i}| = diag(e_i)$. Now, define $\tilde{D}_{2i-1} = D_{1,i}, \tilde{D}_{2i} = D_{2,i}$ for $i \in [n]$, and choose u_i vectors such that

$$(X u_i)[i, 1] = y[i, 1]$$

where the notation $A[i, j]$ denotes the ith row, jth column of A. We can always choose such a vector, as we can choose arbitrary vector v and scale c_i until it matches $c_i(Xv)[i, 1] = y[i, 1]$. Hence, we have found such $\tilde{D}$s that we proposed earlier. □

Proposition B.2. *(Proposition 3.2. of the paper) Suppose we sampled*

$$\tilde{P} = 2\kappa \log(\frac{n}{\delta})$$

hyperplane arrangement patterns, provided that $\mathcal{M}$ is invertible. Then, with probability at least $1 - \delta$, there exists $u_1, u_2, ..., u_{\tilde{P}}$ satisfying

$$\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i = y.$$

Proof. Think of $X_i = \tilde{D}_i X X^T \tilde{D}_i$ as a random independent series of symmetric real $n \times n$ matrices. Also, we know that for any vector $u \in \mathbb{R}^n$,

$$\sqrt{u^T X_i u} \leq \sqrt{\lambda_{max}(XX^T)} \|\tilde{D}_i u\|_2 \leq \sqrt{\lambda_{max}(XX^T)} \|u\|_2,$$

hence we know that for all X_is, $X_i \preccurlyeq \lambda_{max}(XX^T) I$. Now, from Lemma A.1, take $\delta = 1$. Then, we obtain

$$\mathbb{P}(\lambda_{min}(\sum_{i=1}^{\tilde{P}} X_i) \leq 0) \leq n e^{-\mu_{min}/2\lambda_{max}(XX^T)}$$

where $\mu_{min} = \tilde{P}\lambda_{min}(\mathcal{M})$. Hence, when we plug in

$$\tilde{P} = 2\kappa \log(\frac{n}{\delta}),$$

we can see that $\mathbb{P}(\lambda_{min}(\sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i) \leq 0) \leq \delta$, and $\sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i$ is invertible with probability $1-\delta$.

At last, consider the augmented matrix $\mathcal{X} = [\tilde{D}_1 X | \tilde{D}_2 X | ... | \tilde{D}_{\tilde{P}} X]$, which is a $n \times \tilde{P}d$ matrix. As $\mathcal{X}\mathcal{X}^T = \sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i$ is invertible, we can see that $\mathcal{X}$ has n nonzero singular values, hence has rank n. This means that the column space of $[\tilde{D}_1 X | \tilde{D}_2 X | ... | \tilde{D}_{\tilde{P}} X]$ also has rank n, and we can find vectors $u_1, u_2, ..., u_{\tilde{P}}$ that satisfies

$$\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i = y$$

for any given $y \in \mathbb{R}^n$. □

The two propositions directly lead to the following corollary.

Corollary B.3. *(Corollary 3.3. of the paper) Suppose p_0^* and p_1^* are solutions to the unconstrained problem that uses all possible hyperplane arrangement patterns and its randomized relaxation, respectively, and the regularization $\beta = 0$. Also, suppose $\tilde{P} \geq 2\kappa \log(\frac{n}{\delta})$. Then, $p_0^* = p_1^* = 0$ with probability at least $1-\delta$.*

Proof. From Proposition 3.1, we can find $v_1, v_2, ..., v_{2n}$ that satisfies

$$\sum_{i=1}^{2n} D_i' X v_i = y$$

for specific D_i's. Hence, for the unconstrained problem with all possible hyperplane arrangements, we can choose u_is to be $u_i = v_i$ if $D_i = D_i'$, $u_i = 0$ otherwise to perfectly fit a given vector $y \in \mathbb{R}^n$. Also, from Proposition 3.2, we can fit a given vector y with probabillity at least $1-\delta$ when we sample hyperplane arrangement patterns more than $2\kappa \log(\frac{n}{\delta})$ times. Hence, with probability at least $1-\delta$, we know that $p_1^* = 0$. □

Theorem B.4. *(Theorem 3.4. of the paper) Suppose*

$$p^* = \min_{u_i \in \mathbb{R}^d} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y \|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} \|u_i\|_2^2. \tag{11}$$

Then, with probability at least $1-\delta$ there exists scalars $C_1, C_2 > 0$ that satisfies

$$\frac{C_1}{\tilde{P}} \leq p^* \leq \frac{C_2}{\tilde{P}},$$

provided that $\tilde{P} \geq 12\kappa \log(\frac{2n}{\delta})$ and $\mathcal{M}$ is invertible.

Proof. Solving the L_2 regularized problem (11) is equivalent to solving

$$\min_{u_i \in \mathbb{R}^d, w} \frac{1}{2} \|w - y\|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} \|u_i\|_2^2. \tag{12}$$

subject to

$$w = \sum_{i=1}^{\tilde{P}} D_i X u_i.$$

The lagrangian of problem (12) becomes

$$L(u, w, \lambda) = \frac{1}{2} \|w - y\|_2^2 + \lambda^T (w - \sum_{i=1}^{\tilde{P}} D_i X u_i) + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} \|u_i\|_2^2.$$

As the constraint is linear equality and the objective is convex with respect to the arguments, strong duality holds and

$$p^* = \min_{u_i \in \mathbb{R}^d, w \in \mathbb{R}^n} \max_{\lambda \in \mathbb{R}^n} L(u_i, w, \lambda) = \max_{\lambda \in \mathbb{R}^n} \min_{u_i \in \mathbb{R}^d, w \in \mathbb{R}^n} L(u_i, w, \lambda).$$

For a given λ, we can optimize for u_i and w to minimize $L(u_i, w, \lambda)$, where we get $w = y - \lambda$ and $u_i = \frac{1}{\beta} X^T \tilde{D}_i \lambda$. Substituting leads to

$$p^* = \max_{\lambda \in \mathbb{R}^n} -\frac{1}{2}\|\lambda\|_2^2 + \lambda^T y - \frac{1}{2\beta}\lambda^T \big(\sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i\big)\lambda.$$

Now, let's write $\frac{1}{\tilde{P}} \sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i = \mathcal{M}_{\tilde{P}}$. Then, the dual problem can be simplified as

$$p^* = \max_{\lambda \in \mathbb{R}^n} -\frac{1}{2\beta}\lambda^T(\beta I + \tilde{P}\mathcal{M}_{\tilde{P}})\lambda + \lambda^T y$$

and the optimum p^* is given as

$$p^* = \frac{\beta}{2} y^T (\beta I + \tilde{P}\mathcal{M}_{\tilde{P}})^{-1} y.$$

When $\tilde{P} \geq 12\kappa \log(\frac{2n}{\delta})$, by Lemma A.1 we can see that

$$\mathbb{P}\left(\lambda_{min}(\mathcal{M}_{\tilde{P}}) \geq \frac{\lambda_{min}(\mathcal{M})}{2}\right) \geq 1 - \frac{\delta}{2}$$

and

$$\mathbb{P}\left(\lambda_{max}(\mathcal{M}_{\tilde{P}}) \leq \frac{3\lambda_{max}(\mathcal{M})}{2}\right) \geq 1 - \frac{\delta}{2}.$$

Hence, with probability at least $1 - \delta$, we can see that

$$(\beta + \frac{\tilde{P}}{2}\lambda_{\min}(\mathcal{M}))I \preccurlyeq \tilde{P}\mathcal{M}_{\tilde{P}} + \beta I \preccurlyeq (\beta + \frac{3\tilde{P}}{2}\lambda_{\max}(\mathcal{M}))I$$

and we can find $\mathcal{R}_1, \mathcal{R}_2 > 0$ that satisfies

$$\mathcal{R}_1 \tilde{P} I \preccurlyeq \tilde{P}\mathcal{M}_{\tilde{P}} + \beta I \preccurlyeq \mathcal{R}_2 \tilde{P} I.$$

An example is $\mathcal{R}_1 = \frac{\lambda_{\min}(\mathcal{M})}{2}$, $\mathcal{R}_2 = \frac{3\lambda_{\min}(\mathcal{M})}{2} + 1$, provided that $\tilde{P}$ is larger than β. This means that

$$\frac{\beta\|y\|_2^2}{2\mathcal{R}_2\tilde{P}} \leq p^* \leq \frac{\beta\|y\|_2^2}{2\mathcal{R}_1\tilde{P}},$$

and take $C_1 = \frac{\beta\|y\|_2^2}{2\mathcal{R}_2}, C_2 = \frac{\beta\|y\|_2^2}{2\mathcal{R}_1}$ to finish the proof. □

Proposition B.5. *(Proposition 3.5. of the paper) Suppose*

$$p_1^* = \min_{u_i \in \mathbb{R}^d} \frac{1}{2}\|\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y\|_2^2 + \beta \sum_{i=1}^{\tilde{P}} \|u_i\|_2,$$

and $\tilde{P} \geq 8\kappa \log(\frac{n}{\delta})$. *Then with probability at least* $1 - \delta$,

$$p_1^* \leq \sqrt{2}\beta \frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}},$$

provided that $\mathcal{M}$ *is invertible.*

Proof. We know that

$$\begin{aligned} p_1^* &= \min_{u_i \in \mathbb{R}^d} \frac{1}{2} \|\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y\|_2^2 + \beta \sum_{i=1}^{\tilde{P}} \|u_i\|_2 \\ &= \min_{\kappa_i \in \mathbb{R}, u_i \in \mathbb{R}^d} \frac{1}{2} \|\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y\|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} (\kappa_i \|u_i\|_2^2 + \frac{1}{\kappa_i}) \\ &= \min_{\kappa_i \in \mathbb{R}} \min_{u_i \in \mathbb{R}^d} \frac{1}{2} \|\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y\|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} (\kappa_i \|u_i\|_2^2 + \frac{1}{\kappa_i}). \end{aligned}$$

With the same idea from the proof of Equation (11), we can see that when κ_is are given, the dual problem becomes

$$\max_{\lambda \in \mathbb{R}^n} -\frac{1}{2\beta} \lambda^T (\beta I + \sum_{i=1}^{\tilde{P}} \frac{1}{\kappa_i} \tilde{D}_i X X^T \tilde{D}_i) \lambda + \lambda y,$$

and the inner minimization problem has the minimum

$$\min_{u_i \in \mathbb{R}^d} \frac{1}{2} \|\sum_{i=1}^{\tilde{P}} \tilde{D}_i X u_i - y\|_2^2 + \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} (\kappa_i \|u_i\|_2^2 + \frac{1}{\kappa_i}) = \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} \frac{1}{\kappa_i} + \frac{\beta}{2} y^T (\beta I + \sum_{i=1}^{\tilde{P}} \frac{1}{\kappa_i} \tilde{D}_i X X^T \tilde{D}_i)^{-1} y.$$

Hence, we can see that

$$\begin{aligned} p_1^* &= \min_{\kappa_i \in \mathbb{R}} \frac{\beta}{2} \sum_{i=1}^{\tilde{P}} \frac{1}{\kappa_i} + \frac{\beta}{2} y^T (\beta I + \sum_{i=1}^{\tilde{P}} \frac{1}{\kappa_i} \tilde{D}_i X X^T \tilde{D}_i)^{-1} y \\ &\leq \frac{\beta}{2} \min_{\kappa \in \mathbb{R}} y^T (\beta I + \frac{\kappa}{\tilde{P}} \sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i)^{-1} y + \kappa \\ &\leq \frac{\beta}{2} \min_{\kappa \in \mathbb{R}} \frac{y^T (\frac{1}{\tilde{P}} \sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i)^{-1} y}{\kappa} + \kappa \\ &\leq \beta \sqrt{y^T (\frac{1}{\tilde{P}} \sum_{i=1}^{\tilde{P}} D_i X X^T D_i)^{-1} y} \\ &\leq \sqrt{2} \beta \frac{\|y\|_2}{\lambda_{\min}(\mathcal{M})} \end{aligned}$$

Where the last inequality follows from matrix Chernoff that with probability at least $1 - \delta$, the minimum eigenvalue

$$\lambda_{\min}(\frac{1}{\tilde{P}} \sum_{i=1}^{\tilde{P}} \tilde{D}_i X X^T \tilde{D}_i) \geq \frac{\lambda_{\min}(\mathcal{M})}{2}.$$

□

Proposition B.6. *(Proposition 3.6. of the paper) Suppose*

$$p_0^* = \min_{u_i \in \mathbb{R}^d} \frac{1}{2} \|\sum_{i=1}^{P} D_i X u_i - y\|_2^2 + \beta \sum_{i=1}^{P} \|u_i\|_2.$$

Here D_is are all possible hyperplane arrangement patterns. Also, write $\mathcal{M}_i = D_i X X^T D_i$. Then,

$$G\beta \frac{\|y\|_2}{\sqrt{\lambda_{\max}(X X^T)}} \leq p_0^*,$$

where

$$G = 1 - \frac{\beta}{2 \max_{i \in [P]} \sqrt{y^T \mathcal{M}_i y}}$$

Proof. Let's think of the dual problem of the original optimization problem. The lagrangian $L(u_i, w, \lambda)$ is given as

$$L(u_i, w, \lambda) = \frac{1}{2}\|w - y\|_2^2 + \beta \sum_{i=1}^{P} \|u_i\|_2 + \lambda^T (w - \sum_{i=1}^{P} D_i X u_i),$$

and the dual problem is $\max_\lambda \min_{u_i, w} L(u_i, w, \lambda)$. As there is only a linear equality constraint and strong duality holds, we can see that

$$p_0^* = \max_\lambda \min_{u_i, w} \frac{1}{2}\|w - y\|_2^2 + \beta \sum_{i=1}^{P} \|u_i\|_2 + \lambda^T (w - \sum_{i=1}^{P} D_i X u_i).$$

When λ is fixed, the inner minimization problem can be solved as:

i) If $\|X^T D_i \lambda\|_2 > \beta$ for some i: Take $u_i = C X^T D_i \lambda$ and send C to infinity to obtain $-\infty$ as the solution to the inner minimization problem.

ii) If $\|X^T D_i \lambda\|_2 \leq \beta$ for all i: Each $u_i = 0$ when the inner minimization problem is solved, due to Cauchy-Schwartz and

$$\beta \|u_i\|_2 \geq \|X^T D_i \lambda\|_2 \|u_i\|_2 \geq \lambda^T D_i X u_i$$

holds for all $u_i \in \mathbb{R}^d$. Also, $w = y - \lambda$ should hold, and the objective becomes maximizing $-\frac{1}{2}\|\lambda\|_2^2 + \lambda^T y$.

From i), ii), we can see that the dual problem becomes

$$\max_\lambda -\frac{1}{2}\|\lambda\|_2^2 + \lambda^T y$$

subject to

$$\|X^T D_i \lambda\|_2 \leq \beta \quad \forall i \in [P].$$

Now, let's find the maximal scaling coefficient k and the corresponding $\lambda_k = ky$ that meets all the constraints. For that λ, it is clear that $p_0^* \geq -\frac{1}{2}\|\lambda_k\|_2^2 + \lambda_k^T y$. When we substitute ky to λ, we get the constraint of k for each i:

$$k\|X^T D_i y\|_2 \leq \beta, \quad k \leq \frac{\beta}{\sqrt{y^T D_i X X^T D_i y}}$$

must hold. When we write $D_i X X^T D_i = \mathcal{M}_i$, k should satisfy

$$k \leq \frac{\beta}{\sqrt{y^T \mathcal{M}_i y}}$$

for all $i \in [\tilde{P}]$. Choose $k = \frac{\beta}{\max_i \sqrt{y^T \mathcal{M}_i y}}$. Substituting ky to the dual problem leads

$$p_0^* \geq -\frac{\beta^2 \|y\|_2^2}{2 \max_i \{y^T \mathcal{M}_i y\}} + \frac{\beta \|y\|_2^2}{\max_i \{\sqrt{y^T \mathcal{M}_i y}\}} = G \frac{\beta \|y\|_2^2}{\max_i \{\sqrt{y^T \mathcal{M}_i y}\}}.$$

At last, we know that $\mathcal{M}_i \preccurlyeq \lambda_{\max}(XX^T) I$ for all $i \in [P]$. Hence, $y^T \mathcal{M}_i y \leq \lambda_{\max}(XX^T)\|y\|_2^2$ for all $i \in [P]$ and

$$p_0^* \geq G\beta \frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}}.$$

□

Corollary B.7. *(Corollary 3.7. of the paper) Suppose further that* $\|y\|_2 \sqrt{\lambda_{\min}(\mathcal{M})} \geq \beta$. *Then,*

$$\frac{\beta}{2} \frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}} \leq p_1^*.$$

Proof. We know that $G = 1 - \frac{\beta}{2 \max_{i \in [P]} \sqrt{y^T \mathcal{M}_i y}} \geq 1 - \frac{\beta}{2\|y\|_2 \sqrt{\lambda_{\min}(\mathcal{M})}} \geq \frac{1}{2}$. □

Theorem B.8. *(Theorem 3.8. of the paper) Let p_1^* and $\tilde{p}_1^*$ be optimal values of problem (5) with all possible hyperplane arrangements and randomly sampled arrangements, respectively. Suppose we sampled $\tilde{P} \geq 8\kappa \log(\frac{n}{\delta})$ hyperplane arrangement patterns and $\mathcal{M}$ is invertible. We have*

$$0 \leq \tilde{p}_1^* - p_1^* \leq \sqrt{2}\beta \frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}},$$

and

$$p_1^* \leq \tilde{p}_1^* \leq \frac{\sqrt{2\kappa}}{G} p_1^*,$$

with probability at least $1 - \delta$.

Proof. The two inequalities directly follow from Proposition 3.5 and Proposition 3.6 □

Proposition B.9. *$\mathcal{M}_{ij} = \{\frac{1}{2} - \frac{1}{2\pi}\arccos(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2})\}(x_i \cdot x_j)$ for all $i, j \in [n]$. Here, $x \cdot y$ denotes the innter product between x and y.*

Proof. First, recall that $\mathcal{M} = \mathbb{E}_{g\sim N(0,I_d)}[diag[\mathbb{1}(Xg \geq 0)]XX^T diag[\mathbb{1}(Xg \geq 0)]]$. Hence, $\mathcal{M}_{ij}$ has the expression

$$\mathcal{M}_{ij} = \mathbb{P}(x_i \cdot g \geq 0)\mathbb{P}(x_j \cdot g \geq 0)(x_i \cdot x_j).$$

We know that for fixed x_i, x_j, the orthant probability is given as

$$\mathbb{P}(x_i \cdot g \geq 0)\mathbb{P}(x_j \cdot g \geq 0) = \frac{1}{2} - \frac{1}{2\pi}\arccos(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2}),$$

where $g \sim \mathcal{N}(0, I_d)$. This directly implies the claim. □

Theorem B.10. *(Theorem 3.9. of the paper) Let each row x_i of X is sampled i.i.d. from $N(0, I_d)$. Furthermore, suppose $c, \delta > 0$ are given and $n/d = c$ is fixed. There exists d_1 such that if $d \geq d_1$, with probability at least $1 - \delta'$, we have*

$$\lambda_{min}(\mathcal{M}) \geq \frac{d}{10}.$$

Proof. Note that $f(t) = (\frac{1}{2} - \frac{1}{2\pi}\arccos(t))t$ is not differentiable at a neighborhood around $t = 1$. Hence, we need $\tilde{f}_\epsilon$, a C^1 approximator of f, defined as

$$\tilde{f}_\epsilon(x) = \begin{cases} f(x) \quad when \quad x \leq 1 - \epsilon \\ f'(1-\epsilon)x + f(1-\epsilon) - (1-\epsilon)f'(1-\epsilon) \quad when \quad x \geq 1 - \epsilon. \end{cases}$$

Let $\tilde{M}$ be a matrix that satisfies

$$[\tilde{M}]_{ij} = \tilde{f}_{\epsilon_0}(\frac{x_i \cdot x_j}{d}),$$

where $\epsilon_0 = \frac{1}{100c} \leq \frac{1}{100}$. From Lemma A.4, there exists d_2 such that for every $d \geq d_2$,

$$\lambda_{min}(\tilde{M}) \geq \lambda_{min}(\tilde{M}') - \epsilon_0.$$

Now, there exists d_3 such that when $d \geq d_3$, we can decompose $\frac{\mathcal{M}}{d}$ as

$$\frac{\mathcal{M}}{d} = \tilde{M} + diag(\frac{\|x_i\|_2^2}{2d} - \tilde{f}_{\epsilon_0}(\frac{\|x_i\|_2^2}{d})) + \mathcal{M}_e,$$

where

$$[\mathcal{M}_e]_{ij} = \begin{cases} 0 \quad if \quad i = j \\ \frac{1}{2\pi}(-\arccos(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2}) + \arccos(\frac{x_i \cdot x_j}{d}))\frac{x_i \cdot x_j}{d} \quad if \quad i \neq j. \end{cases}$$

with probability at least $1-\delta/2$. Such d_3 exists as for each $i \neq j$, $\frac{x_i \cdot x_j}{d} \sim \mathcal{N}(0, \frac{1}{d})$ and

$$\mathbb{P}(\frac{|x_i \cdot x_j|}{d} \geq \frac{1}{2}) \leq 2\exp(-\frac{d}{8}),$$

meaning that for all $i \neq j$,

$$\mathbb{P}(\max_{i \neq j} \frac{|x_i \cdot x_j|}{d} \leq \frac{1}{2}) \geq 1 - 2n^2\exp(-\frac{d}{8}) = 1 - 2c^2d^2\exp(-\frac{d}{8}),$$

and choose d_3 large enough so that $2c^2d^2\exp(-\frac{d}{8}) \leq \frac{\delta}{2}$. This means with probability at least $1-\delta/2$, $f(\frac{x_i \cdot x_j}{d}) = \tilde{f}_{\epsilon_0}(\frac{x_i \cdot x_j}{d})$ for all $i \neq j$, meaning that the off-diagonal entries of $\frac{\mathcal{M}}{d}$ can be decomposed to $\tilde{M}$ and $\mathcal{M}_e$. Now, Weyl's inequality leads to

$$\frac{\lambda_{\min}(\mathcal{M})}{d} \geq \lambda_{\min}(\tilde{\mathcal{M}}) + \min(\frac{\|x_i\|_2^2}{2d}) - \max(\tilde{f}_{\epsilon_0}(\frac{\|x_i\|_2^2}{d})) - \|\mathcal{M}_e\|_F. \tag{13}$$

Then, from the concentration inequality of Chi-square random variables, we know that there exists d_4 such that for all $d \geq d_4$,

$$(1-\epsilon_0)d \leq (1 - \frac{\epsilon_0}{\log 2n})d \leq \|x_i\|_2^2 \leq (1 + \frac{\epsilon_0}{\log 2n})d \leq (1+\epsilon_0)d, \quad \forall \quad i \in [n], \tag{14}$$

with probability at least $1-\delta/2$. Hence, when $d \geq \max\{d_2, d_3, d_4\}$, both inequalities (13) and (14) satisfy with probability at least $1-\delta$. Use (13) and (14) to get

$$\begin{aligned}
\frac{\lambda_{\min}(\mathcal{M})}{d} &\geq \lambda_{\min}(\tilde{\mathcal{M}}) + \min(\frac{\|x_i\|_2^2}{2d}) - \max(\tilde{f}_{\epsilon_0}(\frac{\|x_i\|_2^2}{d})) - \|\mathcal{M}_e\|_F \\
&\geq \lambda_{\min}(\tilde{\mathcal{M}}) + \frac{1}{2}(1-\epsilon_0) - \tilde{f}_{\epsilon_0}(1+\epsilon_0) - \|\mathcal{M}_e\|_F \\
&\geq \lambda_{\min}(\tilde{\mathcal{M}}') - \epsilon_0 + \frac{1}{2}(1-\epsilon_0) - \tilde{f}_{\epsilon_0}(1+\epsilon_0) - \|\mathcal{M}_e\|_F.
\end{aligned}$$

The last inequality follows from the fact that $d \geq d_2$. We also know that

$$\lambda_{\min}(\tilde{M}') \geq (\tilde{f}_{\epsilon_0}(1) - \tilde{f}_{\epsilon_0}(0) - \tilde{f}'_{\epsilon_0}(0)),$$

as both $11^T, XX^T$ are not invertible. Moreover, $\tilde{f}_{\epsilon_0}(0) = f(0) = 0$, $\tilde{f}'_{\epsilon_0}(0) = f'(0) = 1/4$. Substitute to get

$$\begin{aligned}
\frac{\lambda_{\min}(\mathcal{M})}{d} &\geq \tilde{f}_{\epsilon_0}(1) - \frac{1}{4} - \epsilon_0 + \frac{1}{2}(1-\epsilon_0) - \tilde{f}_{\epsilon_0}(1+\epsilon_0) - \|\mathcal{M}_e\|_F. \\
&\geq \frac{1}{4} - \frac{3}{2}\epsilon_0 - \epsilon_0 f'(1-\epsilon_0) - \|\mathcal{M}_e\|_F \\
&\geq 0.209 - \|\mathcal{M}_e\|_F.
\end{aligned}$$

At last, upper bounding $\|\mathcal{M}_e\|_F$ would be enough to prove the claim. First, we know that when $|u|, |v| \leq \frac{2}{3}$, we know that

$$|\arccos(u) - \arccos(v)| \leq |\arccos'(2/3)||u-v| \leq 2|u-v|.$$

This leads to

$$\begin{aligned}
\|\mathcal{M}_e\|_F &\leq \frac{1}{2\pi}\sqrt{\sum_{i=1}^n \sum_{j \neq i} |-\arccos(\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2}) + \arccos(\frac{x_i \cdot x_j}{d})|^2 |\frac{x_i \cdot x_j}{d}|^2} \\
&\leq \frac{2}{2\pi}\sqrt{\sum_{i=1}^n \sum_{j \neq i} |-\frac{x_i \cdot x_j}{\|x_i\|_2\|x_j\|_2} + \frac{x_i \cdot x_j}{d}|^2 |\frac{x_i \cdot x_j}{d}|^2} \\
&\leq \frac{\epsilon_0}{\pi(1-\epsilon_0)\log 2n}\sqrt{\sum_{i=1}^n \sum_{j \neq i} |\frac{x_i \cdot x_j}{d}|^4}
\end{aligned}$$

with probability at least $1-\delta$. We know from the concentration inequality of maximum of absolute values that

$$\mathbb{P}(\max_{i\leq j}|x_i\cdot x_j|\leq 4\sqrt{2\log 2n}\|x_i\|_2)\geq 1-\frac{2}{(2n)^9},\quad \forall\quad i\in[n]. \tag{15}$$

Hence, with probability at least $1-\delta-\frac{1}{(2n)^8}$, we know that both (14) and

$$\max_{i\leq j}|x_i\cdot x_j|\leq 4\sqrt{2\log 2n\cdot(1+\epsilon_0)d}$$

holds. Hence,

$$\begin{aligned}
\|\mathcal{M}_e\|_F &\leq \frac{\epsilon_0}{\pi(1-\epsilon_0)\log 2n}\sqrt{\sum_{i=1}^n\sum_{j\neq i}|\frac{x_i\cdot x_j}{d}|^4}\\
&\leq \frac{\epsilon_0}{\pi(1-\epsilon_0)d^2\log 2n}\sqrt{\sum_{i=1}^n\sum_{j\neq i}|x_i\cdot x_j|^4}\\
&\leq \frac{\epsilon_0 n}{\pi(1-\epsilon_0)d^2\log 2n}\cdot 32\log 2n(1+\epsilon_0)d\\
&\leq \frac{32\epsilon_0(1+\epsilon_0)c}{\pi(1-\epsilon_0)}\\
&\leq \frac{32}{100\pi}\frac{101}{99}\leq 0.104
\end{aligned}$$

Hence, we know that with probability at least $1-\delta'$,

$$\lambda_{\min}(\mathcal{M})\geq\frac{d}{10}$$

holds, provided that $d\geq d_2,d_3,d_4$. Choose $d_1\geq\max\{d_2,d_3,d_4\}$ sufficiently large. □

Corollary B.11. *(Corollary 3.10. of the paper) When the conditions of Theorem 3.9 holds, with probability at least $1-\delta'-e^{-Cn}$, we have*

$$\kappa\leq 20(\sqrt{c}+1)^2,$$

for some $C>0$. Moreover, let p_1^* and $\tilde{p}_1^*$ be optimal values of problem (5) with all possible hyperplane arrangements and randomly sampled arrangements, respectively. When we sample $\tilde{P}\geq 160(\sqrt{c}+1)^2\log(\frac{n}{\delta})$ hyperplane arrangement patterns, we have

$$p_1^*\leq\tilde{p}_1^*\leq\frac{2\sqrt{10}}{G}(\sqrt{c}+1)p_1^*$$

with probability at least $1-\delta-\delta'-e^{-Cn}$.

Proof. We know that

$$\sqrt{\lambda_{max}(XX^T)}\leq\sqrt{2}(\sqrt{c}+1)\sqrt{d},$$

holds with a high probability that decays exponentially, and we may write it holds with probability at least $1-e^{-Cn}$ for some positive constant C. Also, from Theorem 3.9, we know that

$$\sqrt{\lambda_{min}(\mathcal{M})}\geq\frac{1}{\sqrt{10}}\sqrt{d}.$$

with probability at least $1-\delta'$. Combine the two results to obtain the wanted upper bound on κ. Also, we know that with probability at least $1-\delta$,

$$p_1^*\leq\tilde{p}_1^*\leq\frac{\sqrt{2\kappa}}{G}p_1^*,$$

from Theorem 3.8. Directly using the upper bound on κ leads to the following result. □

C. Proofs in Section 4.

Proposition C.1. *(Proposition 4.2. of the paper) For subsampled hyperplane arrangement patterns $\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_{\tilde{P}}$, let*

$$p_0^* = \min_{u_i, v_i \in \mathcal{K}_{\tilde{D}_i}} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X (u_i - v_i) - y \|_2^2 + \beta \sum_{i=1}^{\tilde{P}} (\|u_i\|_2 + \|v_i\|_2), \tag{16}$$

$$p_1^* = \min_{w_i \in \mathbb{R}^d} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X w_i - y \|_2^2 + \beta \sum_{i=1}^{\tilde{P}} \|w_i\|_2, \tag{17}$$

and $\mathcal{K}_{\tilde{D}_i} = \{u | (2\tilde{D}_i - I) X u \geq 0\}$. Suppose problem (17) has solutions w_i^* for $i \in [\tilde{P}]$, and let

$$C = \max_{i \in [\tilde{P}]} C(\mathcal{K}_{\tilde{D}_i}, \frac{w_i^*}{\|w_i^*\|_2}).$$

Then, $p_0^* \leq C p_1^*$ holds.

Proof. For the optimum w_i^* of problem (17), decompose w_i^* into u_i^*, v_i^* such that $\|u_i^*\|_2 + \|v_i^*\|_2$ is minimal, $u_i^*, v_i^* \in K_i$ and $u_i^* - v_i^* = w_i^*$. Then,

$$\frac{\|u_i^*\|_2 + \|v_i^*\|_2}{\|w_i^*\|_2} \leq C(\mathcal{K}_{\tilde{D}_i}, \frac{w_i^*}{\|w_i^*\|_2}) \leq C$$

and when we substitute u_i^*, v_i^* in (2), we can see that the result would not be greater than $C p_1^*$, as the regression loss is identical and the regularization loss does not blow up C times. Hence, we may conclude that $p_0^* \leq C p_1^*$. □

Proposition C.2. *(Proposition 4.3. in the paper) Take any unit vector z and a cone $\mathcal{K} = \{u | (2D - I) X u \geq 0\}$. If there exists a vector u that satisfies*

$$\|u\|_2 \leq 1, \quad (2D - I) X u \geq \epsilon \cdot |(2D - I) X z|,$$

we know that

$$C(\mathcal{K}, z) \leq 1 + \frac{1}{\epsilon}.$$

Proof. As $(2D - I) X u \geq \epsilon \cdot |(2D - I) X z| \geq (2D - I) X (\pm \epsilon z)$, we know that the vectors $u + \epsilon z, u - \epsilon z \in \mathcal{K}$. Then, the two vectors

$$v_1 = \frac{1}{2}(\frac{u}{\epsilon} + z), \;\; v_2 = \frac{1}{2}(\frac{u}{\epsilon} - z).$$

become two vectors in $\mathcal{K}$ that satisfy $v_1 - v_2 = z$. Hence, the cone sharpness

$$C(\mathcal{K}, z) \leq \|v_1\|_2 + \|v_2\|_2 \leq 1 + \frac{1}{\epsilon} \|u\|_2 \leq 1 + \frac{1}{\epsilon}.$$

where the first inequality follows from the definition of $C(\mathcal{K}, z)$, the second inequality follows from triangular inequality, and the last follows from the fact that u has norm no greater than 1. □

Theorem C.3. *(Theorem 4.4. in the paper) Let $b \in \mathbb{R}^n$ sampled from a folded normal distribution, and let $X \in \mathbb{R}^{n \times d}$ be a matrix where each entries are sampled from a normal distribution. Consider the random variable*

$$F(X, b) = \max_{\|z\|_2 = 1} \min_{\substack{Xu \geq -Xz - kb \\ Xu \geq Xz - kb \\ k \geq 0}} \|u\|_2 + k,$$

where $u, z \in \mathbb{R}^d, k \in \mathbb{R}$. Then, $\mathbb{P}_{X,b}(F(X, b) \leq 200c\sqrt{c \log 2n}) \geq 1 - 1/n^{10} - e^{-Cd}$ for some $C > 0$.

Proof. First, observe the inner minimization problem is strictly feasible for all $\|z\|_2 = 1$, hence strong duality holds. Now, we may write

$$\begin{aligned}
F(X,b) &= \max_{\|z\|_2=1} \min_{\substack{Xu \geq -Xz - kb \\ Xu \geq Xz - kb \\ k \geq 0}} \|u\|_2 + k \\
&= \max_{\|z\|_2=1} \min_{k\geq 0, u} \max_{\lambda,\mu \geq 0} \|u\|_2 + k + \lambda^T(-Xu - Xz - kb) + \mu^T(-Xu + Xz - kb) \\
&= \max_{\substack{\|z\|_2=1 \\ \lambda,\mu\geq 0}} \min_{k\geq 0, u} \|u\|_2 + k + \lambda^T(-Xu - Xz - kb) + \mu^T(-Xu + Xz - kb) \\
&= \max_{\substack{\|z\|_2=1 \\ \lambda,\mu\geq 0}} \min_{k\geq 0, u} \|u\|_2 - (X^T(\lambda+\mu))^T u + k(1 - b^T(\lambda+\mu)) + (X^T(\mu-\lambda))^T z \\
&= \max_{\substack{\lambda,\mu\geq 0 \\ \|X^T(\lambda+\mu)\|_2 \leq 1 \\ b^T(\lambda+\mu)\leq 1}} \|X^T(\lambda-\mu)\|_2.
\end{aligned}$$

Let's write event E_1 to be:

$$E_1 := \sigma_{\max}(X) \leq 2\sqrt{n}.$$

Clearly $\mathbb{P}(E_1) \geq 1 - e^{-C_1 n}$ for some $C_1 > 0$ by Gordon comparison (Thrampoulidis et al., 2014). Now, we write event E_2 to be:

$$E_2 := \max_{\substack{\nu\geq 0 \\ \|X^T\nu\|_2 \leq 1 \\ b^T\nu \leq 1}} \|\nu\|_2 \leq \frac{100\sqrt{\log 2n}}{\sqrt{d}} c = M_0.$$

We show that $\mathbb{P}_{X,b}(E_2) \geq 1 - \frac{1}{n^{20}} - e^{-C_2 d}$. One simple fact we know is that

$$\min_{\substack{\nu\geq 0 \\ \|\nu\|_2 = M \\ b^T\nu \leq 1}} \|X^T\nu\|_2 \geq 1$$

implies

$$\max_{\substack{\nu\geq 0 \\ \|X^T\nu\|_2 \leq 1 \\ b^T\nu \leq 1}} \|\nu\| \leq M.$$

The reason is because if there exists ν^* that satisfies

$$\|\nu\|_2 = M' > M, \nu^* \geq 0, \|X^T\nu^*\|_2 \leq 1, b^T\nu^* \leq 1,$$

we can choose $\frac{M}{M'}\nu^*$ to find a vector that satisfies $\nu \geq 0, \|\nu\|_2 = M, b^T\nu \leq 1$ and $\|X^T\nu\|_2 \leq 1$, which is a contradiction. This means when we define event E_3 to be

$$E_3 := \min_{\substack{\nu\geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} \|X^T\nu\|_2 \geq 1,$$

we know $\mathbb{P}_{X,b}(E_2) \geq \mathbb{P}_{X,b}(E_3)$. We may write the optimization problem in E_3 as

$$\min_{\substack{\nu\geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} \max_{\|\eta\|_2=1} \eta^T X^T \nu.$$

From Gordon comparison, when we define event E_4 to be

$$E_4 := \min_{\substack{\nu\geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} \max_{\|\eta\|_2=1} \|\eta\|_2 g^T\nu + \|\nu\|_2 h^T\eta \geq 1,$$

where $g \sim \mathcal{N}(0, I_n), h \sim \mathcal{N}(0, I_d)$, we have that

$$2\mathbb{P}_{g,h,b}(E_4) \leq \mathbb{P}_{X,b}(E_3) + 1,$$

i.e. an almost sure lower bound of the optimization problem in E_4 acts as an almost sure lower bound of the optimization problem in E_3. Now, define E_5 to be:

$$E_5 := \|h\|_2 \geq \sqrt{d}(1-\epsilon_0), \|g[i_1, i_2, \cdots i_d]_-\|_2 \leq \sqrt{\frac{d}{2}}(1+\epsilon_0),$$
$$\|g\|_\infty \leq 10\sqrt{\log(2n)}, b_{(d+1)} \geq \frac{1}{2c},$$

where $b_{(1)} \leq b_{(2)} \leq \cdots b_{(n)}$ are the order statistics of b, and $b_{(m)} = b_{i_m}$ for $m = 1, 2, \cdots n$, hence i_m indexes the m - th order statistics. Also, ϵ_0 is a fixed constant that can be taken small, and $v_- = v - v_+$ denotes the negative part of v. At last, $g[a_1, a_2, \cdots, a_k]$ denotes the k chosen entries of g, indexed with a_i. As g, h, b are independent, we know that E_5 holds with high probability, i.e. $\mathbb{P}_{g,h,b}(E_5) \geq 1 - \frac{1}{n^{30}} - e^{-C_5 d}$ for some $C_5 > 0$.

We now prove that E_5 implies E_4. First write the optimization problem in E_4 and solve η to get

$$\min_{\substack{\nu \geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} \max_{\|\eta\|_2 = 1} \|\eta\|_2 g^T\nu + \|\nu\|_2 h^T\eta = \min_{\substack{\nu \geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} g^T\nu + \|\nu\|_2\|h\|_2.$$
$$\geq M_0\sqrt{d}(1-\epsilon_0) + \min_{\substack{\nu \geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} g^T\nu.$$

The last inequality follows from E_5. Next, we solve over ν.

$$\begin{aligned}
\min_{\substack{\nu \geq 0 \\ \|\nu\|_2 = M_0 \\ b^T\nu \leq 1}} g^T\nu &\geq \min_{\substack{\nu \geq 0 \\ \|\nu\|_2 \leq M_0 \\ b^T\nu \leq 1}} g^T\nu \\
&= \min_{\nu} \max_{\alpha,\beta,\gamma \geq 0} g^T\nu + \alpha(b^T\nu - 1) + \beta(\|\nu\|_2 - M_0) - \gamma^T\nu \\
&= \max_{\alpha,\beta,\gamma \geq 0} \min_{\nu} g^T\nu + \alpha(b^T\nu - 1) + \beta(\|\nu\|_2 - M_0) - \gamma^T\nu \\
&\geq \max_{\substack{\alpha,\beta,\gamma \geq 0 \\ \|g + \alpha b - \gamma\|_2 \leq \beta}} -(\alpha + M_0\beta) \\
&= \max_{\alpha,\gamma \geq 0} -(\alpha + M_0\|g + \alpha b - \gamma\|_2).
\end{aligned}$$

Now, choose $\alpha_0 = 20c\sqrt{\log(2n)}$, $\gamma_0 = (g + \alpha_0 b)_+$. Then, note that $g + \alpha_0 b$ has positive entries for $i \neq i_1, i_2, \cdots i_d$. That is because assuming E_5, $g_i \geq -10\sqrt{\log 2n}$ and $b_i \geq \frac{1}{2c}$ for $i \neq i_1, i_2, \cdots i_d$. Hence,

$$\|(g + \alpha_0 b)_-\|_2 = \|(g[i_1, i_2, \cdots i_d] + \alpha_0 b[i_1, i_2, \cdots i_d])_-\|_2$$

holds. When we substitute α_0, γ_0, we get

$$\begin{aligned}
\max_{\alpha,\gamma \geq 0} -(\alpha + M_0\|g + \alpha b - \gamma\|_2) &\geq -\alpha_0 - M_0\|g + \alpha_0 b - \gamma_0\|_2 \\
&= -\alpha_0 - M_0\|(g + \alpha_0 b)_-\|_2 \\
&= -\alpha_0 - M_0\|(g[i_1, i_2, \cdots i_d] + \alpha_0 b[i_1, i_2, \cdots i_d])_-\|_2 \\
&\geq -\alpha_0 - M_0\|(g[i_1, i_2, \cdots i_d])_-\|_2.
\end{aligned}$$

The last inequality follows from the fact that for positive vector p and any vector g, $\|(g + p)_-\|_2 \leq \|g_-\|_2$ because adding the positive term only decreases the absolute value of each negative entry, and does not influence positive entries. Now, from E_5, we have

$$\max_{\alpha,\gamma \geq 0} -(\alpha + M_0\|g + \alpha b - \gamma\|_2) \geq -\alpha_0 - M_0(1+\epsilon_0)\sqrt{\frac{d}{2}},$$

and finally we have

$$\begin{aligned}\min_{\substack{\nu\geq 0\\ \|\nu\|_2=M_0\\ b^T\nu\leq 1}} \max_{\|\eta\|_2=1} \|\eta\|_2 g^T\nu + \|\nu\|_2 h^T\eta &\geq M_0\sqrt{d}(1-\epsilon_0) + \min_{\substack{\nu\geq 0\\ \|\nu\|_2=M_0\\ b^T\nu\leq 1}} g^T\nu. \\ &\geq 100(1-\epsilon_0)c\sqrt{\log 2n} - 20c\sqrt{\log 2n} - 100\frac{1+\epsilon_0}{\sqrt{2}}c\sqrt{\log 2n} \\ &\geq 5c\sqrt{\log 2n} \geq 1.\end{aligned}$$

Hence, we have shown that E_5 implies E_4. We know that

$$2\mathbb{P}_{g,h,b}(E_5) \leq 2\mathbb{P}_{g,h,b}(E_4) \leq 1 + \mathbb{P}_{X,b}(E_3) \leq 1 + \mathbb{P}_{X,b}(E_2),$$

and we obtain $\mathbb{P}_{X,h}(E_2) \geq 1 - \frac{1}{n^{20}} - e^{-C_2 d}$, setting $C_2 = C_5\log(2)$ and noticing $n^{30} \geq 2n^{20}$. This means with high probability over X, b,

$$\max_{\substack{\nu\geq 0\\ \|X^T\nu\|_2\leq 1\\ b^T\nu\leq 1}} \|\nu\|_2 \leq \frac{100\sqrt{\log 2n}}{\sqrt{d}}c.$$

The probability that both E_1 and E_2 will happen is at least $1 - \frac{1}{n^{20}} - e^{-Cd}$ for some $C > 0$. When both happens, $F(X,b) \leq 200c\sqrt{c}\sqrt{\log 2n}$. The reason is, for optimal λ^*, μ^* for the dual problem, we have

$$\|\lambda^* + \mu^*\|_2 \leq \frac{100\sqrt{\log 2n}}{\sqrt{d}}c,$$

and we know $\|\lambda^* - \mu^*\|_2 \leq \|\lambda^* + \mu^*\|_2$ as λ^*, μ^* are positive vectors. Now,

$$\begin{aligned}F(X,b) &= \|X^T(\lambda^* - \mu^*)\|_2. \\ &\leq 2\sqrt{n}\|\lambda^* + \mu^*\|_2 \\ &\leq 2\sqrt{n}\frac{100\sqrt{\log 2n}}{\sqrt{d}}c = 200c\sqrt{c}\sqrt{\log 2n}.\end{aligned}$$

Hence E_1, E_2 implies $F(X,b) \leq 200c\sqrt{c}\sqrt{\log 2n}$, and we obtain the wanted result. □

Corollary C.4. *(Corollary 4.5. of the paper) Suppose n, d are sufficiently large that Equation (8) holds with probability at least $1-\delta''$, for $b \in \mathbb{R}^n$ sampled from a folded normal distribution and $X \in \mathbb{R}^{n\times d-1}$ sampled from a normal distribution. Then, with probability at least $1-\delta''$,*

$$C(\mathcal{K}_{\tilde{D}_i}, z) \leq 2 + 200c\sqrt{c\log 2n},$$

also holds for all unit vectors z.

Proof. We prove for the cone that contains e_1, $\mathcal{K} = \{\bar{X}u \geq 0\}$. Take any unit vector z. We know that

$$\max_{\|z\|_2=1} \min_{\substack{Xu\geq -Xz-kb\\ Xu\geq Xz-kb\\ k\geq 0}} \|u\|_2 + k \leq 200c\sqrt{c\log 2n},$$

with probability at least $1-\delta''$. Hence, with probability $1-\delta''$, there exists $u_0 \in \mathbb{R}^{d-1}$, $k_0 \geq 0$ that satisfies

$$\bar{X}[1:n, 2:d]u_0 + \bar{X}[1:n,1]k_0 \geq |\bar{X}[1:n,2:d]z[2:d]|.$$

and $u_0 + k_0 \leq 200c\sqrt{c\log 2n}$. Here, $X[a:b, c:d]$ denotes the submatrix of row a to b, column c to d. For that u_0, k_0, we know that

$$\bar{X}[1:n,2:d]u_0 + \bar{X}[1:n,1](k_0+1) \geq \bar{X}[1:n,1]|z[1]| + |\bar{X}[1:n,2:d]z[2:d]| \geq |\bar{X}z|.$$

Write $u = \begin{bmatrix} k_0+1 \\ u_0 \end{bmatrix}$ to see that $\bar{X}u \geq |\bar{X}z|$. Also, the norm $\|u\|_2$ is bounded by

$$\|u\|_2 \leq 1 + k_0 + \|u_0\|_2 \leq 1 + 200c\sqrt{c\log 2n}.$$

Choose the center in Proposition 4.3 as $u/(1+k_0+\|u_0\|_2)$ and apply the proposition to obtain the wanted result. □

Corollary C.5. *For subsampled hyperplane arrangement patterns $\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_{\tilde{P}}$, let*

$$p_0^* = \min_{u_i, v_i \in \mathcal{K}_{\tilde{D}_i}} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X (u_i - v_i) - y \|_2^2 + \beta \sum_{i=1}^{\tilde{P}} (\|u_i\|_2 + \|v_i\|_2), \tag{18}$$

$$p_1^* = \min_{w_i \in \mathbb{R}^d} \frac{1}{2} \| \sum_{i=1}^{\tilde{P}} \tilde{D}_i X w_i - y \|_2^2 + \beta \sum_{i=1}^{\tilde{P}} \|w_i\|_2. \tag{19}$$

Furthermore, assume (A1) and n, d to be sufficiently large so that Corollary C.4 holds with probability at least $1 - \delta''$. Then, we have $p_1^* \leq p_0^* \leq (2 + 200c\sqrt{c \log 2n}) p_1^*$ with probability at least $1 - \tilde{P}\delta''$.

Proof. The first inequality follows from triangular inequality, and the second from Proposition 4.2. $\tilde{P}$ appears due to union bound, i.e. all cone sharpness constants should be bounded. □

D. Proof of the Main Theorems

Theorem D.1. *(Theorem 2.1. of the paper) Let the optimal value of the 2-layer ReLU network as*

$$p^* = \min_{u,\alpha} \frac{1}{2}\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(\|u_j\|_2^2 + \|\alpha_j\|_2^2),$$

and the convex optimization problem with random hyperplane arrangement patterns as

$$\tilde{p}^* = \min_{u_i, v_i \in \mathcal{K}_{\tilde{D}_i}} \frac{1}{2}\|\sum_{i=1}^{m/2}\tilde{D}_i X(u_i - v_i) - y\|_2^2 + \beta\sum_{i=1}^{m/2}(\|u_i\|_2 + \|v_i\|_2).$$

Suppose that $m = \kappa \max\{m^*, 320(\sqrt{c}+1)^2\log(\frac{n}{\delta})\}$ for fixed $\kappa \geq 1$, where m^* defined as in Section 2.3. Moreover, assume $n/d = c$ is fixed and the entries of X are i.i.d. $\sim \mathcal{N}(0,1)$. At last, let $d \geq d_3$ so that both Corollary C.5 and Corollary 3.10 holds with probability at least $1 - \delta - \delta' - m\delta''$ and $G < 1/2$. Then,

$$p^* \leq \tilde{p}^* \leq 2\sqrt{20}(\sqrt{c}+1)(2 + 200c\sqrt{c\log 2n})\; p^*,$$

holds with probability at least $1 - \delta - \delta' - m\delta''$.

Proof. Let's denote

$$p_2^* = \min_{u_i, v_i \in \mathcal{K}_{D_i}} \frac{1}{2}\|\sum_{i=1}^{P} D_i X(u_i - v_i) - y\|_2^2 + \beta\sum_{i=1}^{P}(\|u_i\|_2 + \|v_i\|_2),$$

$$p_3^* = \min_{u_i \in \mathbb{R}^d} \frac{1}{2}\|\sum_{i=1}^{P} D_i X u_i - y\|_2^2 + \beta\sum_{i=1}^{P}\|u_i\|_2,$$

where $D_1, D_2, ..., D_P$ are all possible hyperplane arrangement patterns, and

$$p_4^* = \min_{u_i \in \mathbb{R}^d} \frac{1}{2}\|\sum_{i=1}^{m/2} \tilde{D}_i X u_i - y\|_2^2 + \beta\sum_{i=1}^{m/2}\|u_i\|_2,$$

where $\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_{m/2}$ are randomly sampled hyperplane arrangement patterns. First, from preliminaries, we know that $p^* = p_2^*$. It is clear that $p_2^* \leq \tilde{p}^*$, as we use less hyperplane arrangement patterns during approximation. Moreover, we know that $p_3^* \leq p_2^*$, as $\|u_i - v_i\|_2 \leq \|u_i\|_2 + \|v_i\|_2$ and $u_i - v_i$ can represent arbitrary vector in $\mathbb{R}^d$ even with the constraint $u_i, v_i \in \mathcal{K}_{\tilde{D}_i}$. To wrap up, we know that

$$p_3^* \leq p_2^* = p^* \leq \tilde{p}^*.$$

From Corollary C.5 we know that $\tilde{p}^* \leq (2 + 200c\sqrt{c\log 2n})\; p_4^*$, and we also know from Proposition 3.6 that $p_3^* \geq G\beta\frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}}$. At last, from Proposition 3.5, we know that $p_4^* \leq \sqrt{2}\beta\frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}}$. Hence, we know that

$$\begin{aligned}
G\beta\frac{\|y\|_2}{\sqrt{\lambda_{\max}(XX^T)}} \leq p_3^* \leq p_2^* = p^* \leq \tilde{p}^* &\leq (2 + 200c\sqrt{c\log 2n})\; p_4^* \\
&\leq \sqrt{2}\beta\frac{\|y\|_2}{\sqrt{\lambda_{min}(\mathcal{M})}} \cdot (2 + 200c\sqrt{c\log 2n}) \\
&\leq \frac{\sqrt{2}}{G}\sqrt{\frac{\lambda_{\max}(XX^T)}{\lambda_{\min}(\mathcal{M})}} \cdot (2 + 200c\sqrt{c\log 2n})\; p^* \\
&\leq \frac{\sqrt{20}}{G}(\sqrt{c}+1)(2 + 200c\sqrt{c\log 2n})\; p^*,
\end{aligned}$$

which finishes the proof that

$$p^* \leq \tilde{p}^* \leq 2\sqrt{20}(\sqrt{c}+1)(2 + 200c\sqrt{c\log 2n})\; p^*.$$

□

Theorem D.2. *(Theorem 2.3. of the paper) Assume (A1) and suppose that* $m = \kappa \max\{m^*, 320(\sqrt{c}+1)^2 \log(n/\delta)\}$ *for* $\kappa \geq 1$*. Then, there exists a randomized algorithm with* $O(d^3m^3)$ *complexity that solves problem* (1) *within* $O(\sqrt{\log n})$ *relative optimality bound with high probability.*

Proof. Consider the Gaussian relaxation of the convex reformulation with $[m/2]$ hyperplane arrangement patterns. As there are $O(dm)$ variables, we can solve the problem with $O(d^3m^3)$ complexity using standard interior point solvers. Moreover, we have the approximation bound of Theorem 2.1, which leads to the fact that the solved global minima has $O(\sqrt{\log 2n})$ guarantees. At last, we can map the solution $\{(u_i^*, v_i^*)\}_{i=1}^{[m/2]}$ to the parameter space of two-layer neural networks with the mapping

$$u_i^* \to (\frac{u_i^*}{\sqrt{\|u_i^*\|_2}}, \sqrt{\|u_i^*\|_2}), v_i^* \to (\frac{v_i^*}{\sqrt{\|v_i^*\|_2}}, -\sqrt{\|v_i^*\|_2}),$$

to find the parameters of the two-layer neural network that has the same loss function value as the optimal value of the convex problem. Hence, we can find parameters of two-layer neural network that has $O(\sqrt{\log n})$ relative optimality bound in $O(d^3m^3)$ time. □

Theorem D.3. *(Theorem 2.5. of the paper) Consider the training problem* $\min_{u_j, \alpha_j} \mathcal{L}(u, \alpha)$*, where the loss function* $\mathcal{L}$ *is given as*

$$\mathcal{L}(u, \alpha) = \frac{1}{2}\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(\|u_j\|_2^2 + \alpha_j^2).$$

Assume (A1),(A2), d *sufficiently large and* $m = \kappa \max\{m^*, 320(\sqrt{c}+1)^2 \log(\frac{n}{\delta})\}$ *for some fixed* $\kappa \geq 1$ *so that Theorem 2.1 holds with probability at least* $1 - \delta - \delta' - m\delta''$*. For any random initialization* $\{u_i^0, \alpha_i^0\}_{i=1}^m$*, suppose local gradient method converged to a stationary point* $\{u_i', \alpha_i'\}_{i=1}^m$*. Then, with probability at least* $1 - \delta - \delta' - m\delta''$*,*

$$\mathcal{L}(u', \alpha') \leq C\sqrt{\log 2n}\mathcal{L}(u^*, \alpha^*),$$

for some $C \geq 1$*. Here,* $\{u_i^*, \alpha_i^*\}_{i=1}^m$ *is a global optimum of* $\mathcal{L}(u, \alpha)$*.*

Proof. We know that for the stationary point $\{u_i', \alpha_i'\}_{i=1}^m$, we have a corresponding convex optimization problem

$$p^* := \min_{u_i, v_i \in \mathcal{K}_{\tilde{D}_i}} \frac{1}{2}\|\sum_{i=1}^{m}\tilde{D}_i(u_i - v_i) - y\|_2^2 + \beta\sum_{i=1}^{m}\|u_i\|_2 + \|v_i\|_2, \tag{20}$$

where $\tilde{D}_i = \mathbb{1}(Xu_i' \geq 0)$ and the solution mapping of the optimal solution $\{(u_i^*, v_i^*)\}_{i=1}^m$ of the convex problem and the stationary point given as

$$\alpha_i' = sign(\alpha_i')\sqrt{\|u_i^*\|_2}, \quad u_i' = \frac{u_i^*}{\sqrt{\|u_i^*\|_2}}\mathbb{1}(\alpha_i' \geq 0) + \frac{v_i^*}{\sqrt{\|v_i^*\|_2}}\mathbb{1}(\alpha_i' < 0).$$

By (A2), we know that at least half of the random hyperplane arrangement patterns are preserved from $\mathbb{1}(Xu_i \geq 0)$ at initialization. Let the hyperplane arrangement patterns be $\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_{m/2}$ without loss of generality. Now, we know that the optimal solution of the problem

$$p' := \frac{1}{2}\|\sum_{i=1}^{m/2}\tilde{D}_i(u_i - v_i) - y\|_2^2 + \beta\sum_{i=1}^{m/2}\|u_i\|_2 + \|v_i\|_2$$

has $O(\sqrt{\log 2n})$ approximation guarantee with probability at least $1 - \delta - \delta' - m\delta''$, and as it is using less hyperplane arrangement patterns, we know that $p^* \leq p'$. At last, we know that $p^* = \mathcal{L}(u', \alpha')$ from solution mapping. Plugging in $\mathcal{L}(u', \alpha')$ at p^* and using Theorem 2.1 on p' yields the wanted result. □

E. Effect of Regularization

In this section, we demonstrate that regularization may affect the test performance of the model. We do two experiments, one with a synthetic dataset with a hidden planted two-layer network with width 50, and the other with MNIST data where digits 0∼4 are labeled 1 and 5∼9 are labeled -1. We use the SCNN library (Mishkin et al., 2022a) to fit the model with an equivalent convex model with different width and regularization. We can observe that the test performance may differ ∼5% for synthetic data, and ∼10% for MNIST when regularization differs. Hence, a good choice of regularization matters regarding finding a good model.

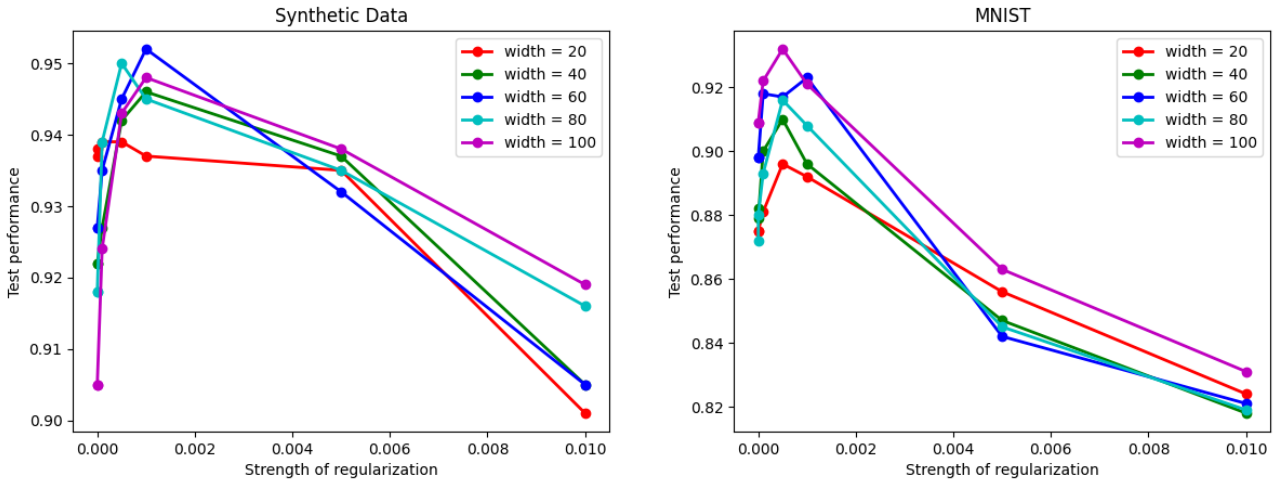

Figure 3. Effect of regularization for test performance. The left and the right figure shows test performance of a trained model for synthetic data and MNIST respectively, for different model sizes and regularization. For different choices of regularization, the test performance changes at maximum 5% for synthetic data, and 10% for MNIST.